

Gutachten im Auftrag des
Bundesministeriums für Gesundheit (BMG)

**Record Linkage
der Daten des FDZ Gesundheit
mit extern gehaltenen Forschungsdaten
anhand des Beispiels einer Verknüpfung
mit Daten einer Kohorte des NAPKON**

Version 2.1
vom 23. Mai 2023

Gutachter

Prof. Dr. med. **Jochen Schmitt**, MPH

Koordination des Gutachtens
Professur für Sozialmedizin und Versorgungsforschung
Zentrum für Evidenzbasierte Gesundheitsversorgung (ZEGV)
Medizinische Fakultät Carl Gustav Carus, TU Dresden sowie
Universitätsklinikum Carl Gustav Carus Dresden

Dr. med. **Peter Ihle**

CIO Datenmanagement, Datenschutz
PMV forschungsgruppe, Klinik und Poliklinik für Psychiatrie,
Psychosomatik und Psychotherapie des Kindes- und Jugendalters
an der Medizinischen Fakultät und Uniklinik Köln, Universität zu Köln

Prof. Dr. med. **Wolfgang Hoffmann**, MPH

Abteilungsleitung Versorgungsepidemiologie und Community Health
Institut für Community Medicine
Universitätsmedizin Greifswald

Prof. Dr. rer. nat. **Jens-Peter Reese**, MPH

Professur für Versorgungsforschung und Public Health
Institut für Klinische Epidemiologie und Biometrie
Julius-Maximilians-Universität Würzburg

Dr. rer. nat. **Olaf Schoffer**, Dipl.-Stat.

Bereichsleiter onkologische Versorgungsforschung
Zentrum für Evidenzbasierte Gesundheitsversorgung (ZEGV)
Medizinische Fakultät Carl Gustav Carus, TU Dresden sowie
Universitätsklinikum Carl Gustav Carus Dresden

Inhaltsverzeichnis

1	Ausgangssituation und Problemanalyse	1
1.1	Umriss des Kontextes	1
1.2	Fallbeispiel: Das Nationale Pandemie Kohorten Netz (NAPKON) des Netzwerk Universitätsmedizin (NUM).....	4
1.3	Machbarkeitsstudie im Fallbeispiel Datenverknüpfung der klinischen NAPKON- POP Daten mit GKV Routinedaten	8
2	Linkage von Forschungsdaten mit GKV-Daten.....	19
2.1	Forschungsbedarfe / Fragestellungen	19
2.2	Datenlinkage	22
2.3	Rechtliche Grundlage	29
2.3.1	Generelle Aspekte der Datennutzung im Kontext des FDZ Gesundheit	29
2.3.2	Rechtliche Aspekte des Datenlinkage von VeDa, insbesondere GKV- Routinedaten.....	31
2.4	Rolle der Krankenkassen	33
2.5	Barrieren: Erfahrungen aus NAPKON.....	35
2.6	Die Rolle des FDZ Gesundheit	38
2.7	Zukünftige Rollen von Krankenkassen und FDZ Gesundheit	41
3	Perspektive und Einordnung der Beteiligten	43
3.1	Sicht der Krankenkassenvertretenden.....	43
3.2	Sicht der Forschung.....	47
3.3	Sicht der Ethik.....	56
3.4	Sicht der Patientenvertretung.....	63
3.5	Sicht der Klinikerinnen und Kliniker (FOSA)	65
4	Empfehlungen	70
4.1	Empfehlungen zum FDZ Gesundheit.....	70
4.1.1	Allgemeine Empfehlungen zum FDZ Gesundheit.....	70
4.1.2	Empfehlungen zum Identitätsmanagement	88
4.1.3	Rechtliche/Juristische Empfehlungen zum FDZ Gesundheit.....	90
4.2	Anwendung auf NAPKON	92
4.3	Weitere Empfehlungen	98
5	Zusammenfassung.....	103
6	Danksagung	110
7	Literaturverzeichnis.....	111
8	Anhang	118

8.1	Anlage 1a: Studienprotokoll der Proof-of-Concept Studie	118
8.2	Anlage 1b: Probandeninformation (NAPKON – POP Kohortenstudie)	134
8.3	Anlage 1c: Einwilligungserklärung (NAPKON – POP Kohortenstudie)	144
8.4	Anlage 2: FOSA – Befragung.....	148
8.5	Anlage 3: Anschreiben und Fragebogen Patientenvertretungen	168

Tabelle 1: Abkürzungsverzeichnis

Abkürzung	Erläuterung
BAS	Bundesamt für Soziale Sicherung
BDSG	Bundesdatenschutzgesetz
BfArM	Bundesinstitut für Arzneimittel und Medizinprodukte
BfDI	Bundesbeauftragter für den Datenschutz und die Informationsfreiheit
BKK	Betriebskrankenkasse
BMBF	Bundesministerium für Bildung und Forschung
DaTraV	Datentransparenzverordnung
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information
DMP	Disease-Management-Programme
DNVF	Deutsches Netzwerk Versorgungsforschung
DSGVO	Datenschutz-Grundverordnung
DSRV	Datenstelle der Rentenversicherung
DZNE	Deutsches Zentrum für Neurodegenerative Erkrankungen
EIRD	Errichtung des Implantateregisters Deutschland
ePA	elektronische Patientenakte
EU	Europäische Union
FAIR	Findable (Auffindbar), Accessible (Zugänglich), Interoperable (Interoperabel) und Reusable (Wiederverwendbar)
FDZ	Forschungsdatenzentrum
FOSA	Fach- und Organspezifische Arbeitsgruppen
GAP	Gastarbeitsplätze
GEP	Gute Epidemiologische Praxis
GKV	Gesetzliche Krankenversicherung
GPD	Gute Praxis Datenlinkage
GPS	Gute Praxis Sekundärdatenanalyse
HAP	hochauflösende Plattform
IAB	Institut für Arbeitsmarkt und Berufsforschung

ID	Identifikator
IQWIG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITSG	Informationstechnische Servicestelle der gesetzlichen Krankenversicherung
KFRG	Krebsfrüherkennungs- und Registergesetz
KiGGS	Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland
KK	Krankenkassen
KKR	Klinischen Krebsregister
KVNR	Krankenversicherternummer
LP	Lieferpseudonyme
MDAT	Medizinische Daten
MII	Medizininformatikinitiative
Morbi-RSA	Morbiditätsorientierter Risikostrukturausgleich
NAKO	Nationale Kohorte
NAPKON	Nationale Pandemie Kohorten Netz
NUKLEUS	NUM Klinische Epidemiologie- und Studienplattform
NUM	Netzwerk Universitätsmedizin
PID	Pseudonym (pseudonymer Identifikator)
PKV	Privaten Krankenversicherung
POP	Populationsbasierte Plattform
PROs, PROMS	patient reported outcome measures
PüP	periodenübergreifendes Pseudonym
QS	Qualitätssicherung
RKI	Robert-Koch-Institut
RVNR	Rentenversicherungsnummer
SGB	Sozialgesetzbuch
SOP	Standard Operating Procedure
SQL	Structured Query Language
SÜP	sektorenübergreifende Plattform
SVR	Sachverständigenrat

THS	Treuhandstelle
TMF	Technologie- und Methodenplattform für die vernetzte medizinische Forschung
UAC	Use-and-Access-Committee
VeDa	Versorgungsnahe Daten
VPN	Virtual Private Network
VST KVNR	Vertrauensstelle Krankenversicherungsnummer
ZEGV	Zentrum für Evidenzbasierte Gesundheitsversorgung

1 Ausgangssituation und Problemanalyse

1.1 Umriss des Kontextes

Versorgungsnahe Daten (VeDa) sind eine wesentliche Ressource für die wissenschaftlich fundierte Beschreibung, Evaluation und Weiterentwicklung der Gesundheitsversorgung und des Gesundheitssystems [1, 2]. VeDa sind dadurch gekennzeichnet, dass sie möglichst nahe die medizinische Routineversorgung abbilden, ohne dass bei der Gesundheitsversorgung spezifische Vorgaben oder Interventionen erfolgen [3]. Durch die Erhebung im Versorgungskontext erlauben VeDa insbesondere die Untersuchung von Fragestellungen zur Zugangs- und Versorgungsgerechtigkeit, zum Ressourcenbedarf, zum Outcome der Gesundheitsversorgung und auch zum Grad der Implementierung und zur Akzeptanz von Versorgungsangeboten und medizinischen Innovationen. Für wissenschaftlich fundierte Aussagen ist dabei u. a. eine hohe Repräsentativität der untersuchten Population wichtig [1]. VeDa erlauben darüber hinaus auch die Bewertung von Interventionseffekten, wobei jedoch aufgrund der fehlenden Randomisierung ein adäquates Studiendesign (Kontrollgruppenbildung, Matching) und die Berücksichtigung möglichst aller bekannter Confounder in der statistischen Analyse entscheidend für die Nutzbarkeit von VeDa etwa für die Nutzenbewertung sind [3].

Hauptquellen von VeDa sind in Deutschland die Abrechnungsdaten der Gesetzlichen Krankenversicherung (GKV-Routinedaten), Daten aus medizinischen Registern, insbesondere den Krebsregistern nach §65c SGB V und aus Kohortenstudien, bspw. des Nationalen Pandemie Kohorten Netzes (NAPKON) [4], der NAKO Gesundheitsstudie [5], der Study of Health in Pomerania (SHIP) [6-8] oder des KiGGS des Robert-Koch-Institutes [9, 10]. Zukünftig kommen Behandlungsdaten aus Klinik- und Praxisinformationssystemen (z. B. Daten der Datenintegrationszentren der Medizininformatik Initiative (MII), Daten der elektronischen Patientenakte (ePA), Daten aus Gesundheits-Apps und digitalen Gesundheitsanwendungen (DiGa), sowie aus anwendungsbegleitenden Datenerhebungen im Zuge der frühen Nutzenbewertung nach §35a SGB V dazu [11].

VeDa spielen für die Versorgungsforschung und damit für die wissenschaftlich fundierte Bewertung und Weiterentwicklung und Verbesserung der Qualität der Gesundheitsversorgung in Deutschland eine zunehmend zentrale Rolle. Gerade im Rahmen der COVID-19 Pandemie erwiesen sich VeDa und dabei insbesondere GKV-Routinedaten als eine ganz wesentliche Quelle aktueller wissenschaftlicher Evidenz zu drängenden, die gesamte Gesellschaft betreffenden Fragestellungen wie etwa der Impfpriorisierung, der Basisinzidenz möglicher Nebenwirkungen der Impfung und dem Long-COVID Risiko von Erwachsenen, Kindern und Jugendlichen in Deutschland [12].

Das Forschungsdatenzentrum Gesundheit (FDZ Gesundheit) am Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) hat für die Gesundheitsforschung mit VeDa eine

überragende Relevanz, weil es Längsschnittdaten von allen rund 70 Millionen GKV-Versicherten in Deutschland sektorübergreifend zur ambulanten und stationären medizinischen Versorgung, sowie zur Arzneimittelversorgung verfügbar machen wird [13]. Damit erfüllt es eine der Anforderungen des Sachverständigenrat (SVR) Gesundheit im Hinblick auf Forschungsinfrastruktur, dass die "Verfügbarkeit von Daten für die Forschung strategisch weiterentwickelt werden" [14]. Trotz der hohen Repräsentativität und Vollständigkeit der dokumentierten Diagnosen, Prozeduren und Arzneimittelverordnungen zu Lasten der GKV fehlen jedoch je nach Fragestellung häufig zusätzliche Informationen, etwa zu Patienten-berichteten Outcomes (PROs), zu wichtigen Confoundern, zur Erkrankungsschwere, zu soziodemographischen Faktoren, zur Diagnostik und Bildgebung, zu Laborergebnissen und Biomarkern und zu Patientenpräferenzen. Für die qualitativ hochwertige wissenschaftliche Beantwortung von versorgungsrelevanten Fragestellungen müssen die GKV-Daten daher häufig mit anderen VeDa (Registerdaten, Klinikdaten) oder/und mit Befragungsdaten und Daten aus Kohortenstudien verknüpft werden (Data Linkage).

Trotz der hohen Relevanz eines Datalinkages von GKV-Routinedaten mit anderen Datenbeständen gibt es in Deutschland bisher nur einige wenige wegweisende Forschungsprojekte, die erfolgreich ein Datenlinkage von GKV-Routinedaten mit anderen Daten umsetzen konnten. Beispiele sind die NORAH-Studie [15] zum Zusammenhang von Verkehrslärm und Morbidität in der Bevölkerung (kardiovaskuläre Ereignisse, psychische Erkrankungen), die LidA-Studie zu Zusammenhängen von Arbeit, Alter, Gesundheit und Erwerbsteilhabe [16] sowie die EcoCare-PIn-Studie [17] zu Langzeiteffekten von Frühgeburtlichkeit und psychosozialen Interventionsmöglichkeiten. Aktuell prominentestes Beispiel in Deutschland für ein Record-Linkage von primären Forschungsdaten mit GKV-Routinedaten ist die NAKO Gesundheitsstudie. Zum Stand 31.10.2022 haben 205.412 Probandinnen und Probanden die erste Untersuchung abgeschlossen, und bereits 89.743 an der 5-Jahres-Follow-up Untersuchung teilgenommen. Im Zuge der modularen Einwilligungserklärung der NAKO wird für die Verknüpfung mit den Krankenversicherungsdaten von jeder/m TeilnehmerIn eine informierte Einwilligung eingeholt. Dabei wird separat gefragt, ob die Teilnehmer a.) einer Verknüpfung der Untersuchungsdaten mit ihren Krankenversicherungsdaten für den zurückliegenden Zeitraum von 5 Jahren vor der Untersuchung im Rahmen der NAKO und b.) für den folgenden Zeitraum von 5 Jahren in die Zukunft hinein. Nach entsprechender Aufklärung hat die große Mehrheit der Teilnehmende in die Verknüpfung ihrer Krankenkassendaten mit den Untersuchungsdaten eingewilligt, 94,2 % für den rückwirkenden Zeitraum, 93,94% in den prospektiven Zeitraum. Die Verhandlungen mit den gesetzlichen Krankenkassen begannen in 2013/2014, mithin vor mehr als 8 Jahren. Bis zum 31.10.2022 lagen für 59 gesetzliche Krankenkassen Genehmigungen nach §75 SGB X vor. Bei 55 gesetzlichen Krankassen wurden bereits Daten für jeweils dort versicherte NAKO Teilnehmende angefordert. Für 105.000 Teilnehmende wurden bereits Daten an das Datenintegrationszentrum der NAKO übermittelt. Zusätzlich wurden bisher von 4 privaten Krankenkassen Daten zu 3100 bei diesen privat versicherten NAKO-Teilnehmende

angefordert. Bis zum Stichtag haben 3 private Versicherungen Daten zu insgesamt 598 Teilnehmende geliefert. Die große Mehrheit der NAKO-Teilnehmende stimmten auch der Verknüpfung ihrer Daten in den Krebsregistern zu (98,29 % in den epidemiologischen und 98,29 % in den Klinischen Krebsregistern der Länder). Zum Stand 31.10.2022 waren entsprechende Datennutzungsanträge bei 6 klinischen und 8 epidemiologischen Krebsregistern gestellt. Davon wurden in 4 klinischen und 6 epidemiologischen Krebsregistern bereits 9.575 Krebsfälle bei 8.520 Teilnehmende der NAKO erfasst [18].

Bisher war die Kooperation mit einer oder mehreren Krankenkassen die einzige Möglichkeit, um derartige Projekte mit personenbezogenem Linkage von GKV-Daten mit anderen Daten umzusetzen. Dieser Weg hat als Einschränkungen einen sehr hohen Aufwand und die fehlende Planbarkeit für Wissenschaftlerinnen und Wissenschaftler, da Krankenkassen einzeln angesprochen werden müssen und kein Anspruch auf deren Beteiligung besteht. In der Praxis konnten aus diesen Gründen nur selten und dann mit sehr hohem Ressourcen- und Zeitaufwand Forschungsprojekte mit (nahezu) kassenübergreifender Repräsentanz erfolgreich durchgeführt werden.

Es besteht daher ein sehr hoher Bedarf, die Daten des FDZ Gesundheit mit anderen Datenbeständen zu verknüpfen. Aus Sicht der Gutachter ist es daher unbedingt erforderlich, die Dateninfrastruktur und die entsprechenden Gesetzesgrundlagen so weiterzuentwickeln, dass ein Linkage von GKV-Daten des FDZ Gesundheit mit anderen Daten auf Basis einer informierten Einwilligung (Informed Consent) der Versicherten möglich ist. Der Broad Consent der Medizin-Informatik Initiative (MII) enthält vor diesem Hintergrund ein Modul zur Einwilligung von Patientinnen und Patienten der Universitätsmedizin zur Nutzung und Verknüpfung von Daten der MII mit GKV-Routinedaten (GKV-Modul). Momentan wird der Broad Consent schrittweise zunächst in den Universitätskliniken ausgerollt. Allein durch dieses Vorhaben wird ein sehr großer Bedarf bestehen. In einer aktuellen Stellungnahme des Wissenschaftsrates werden für 2019 etwa 1,97 Mio. stationäre Fälle in allen Universitätskliniken angegeben. Damit entfallen etwas über 10% aller stationären Fälle in Deutschland auf die Universitätsmedizin. Aktuell nehmen alle Universitätsklinika an der Medizininformatik-Initiative teil. Im Januar 2023 wurde der Broad Consent an 24 Standorten erhoben. Zu diesem Zeitpunkt lagen insgesamt 105.000 BC vor, davon 65.000 inkl. GKV-Modul (62%). Die Einwilligung zum Broad Consent wird von 87,5% der aufgeklärten Patientinnen und Patienten gegeben. Patientinnen und Patienten[19]. Auf der Basis der bisher vorliegenden Zahlen kann somit hochgerechnet werden, dass bei vollständiger Implementierung des BC bei allen stationären Patientinnen und Patienten und Patienten in Universitätsklinika pro Jahr mindestens 1 Mio. Kassenmodule positiv konsentiert werden. Hinzu kommen Konsente aus der wachsenden Anzahl von nichtuniversitären Partnerkrankenhäusern (z. B. den Digitalen Fortschritts-Hubs der MII).

Ziel dieses Gutachtens ist es, vor dem Hintergrund des im Koalitionsvertrag der Bundesregierung angekündigten Gesundheitsdatennutzungsgesetzes konkrete Vorschläge für die technische, rechtliche und infrastrukturelle Umsetzung einer personenindividuellen Datenlinkage von GKV-Routinedaten des FDZ Gesundheit mit anderen Datenbeständen auf Basis der individuellen informierten Einwilligung von Versicherten zu machen. Als Beispiel wurde hierbei die Linkage von GKV-Routinedaten mit Daten der NAPKON-Kohortenstudie "Populationsbasierte Plattform (POP)" des Netzwerks Universitätsmedizin (NUM) gewählt. Zusätzlich geht das Gutachten auf die zukünftige Rolle der Krankenkassen in Studien der Versorgungsforschung mit Datenlinkage ein, zeigt auf, in welchen Szenarien eine Einbeziehung von Krankenkassen aus Sicht der Forscher weiterhin notwendig ist und gibt Empfehlungen für die Weiterentwicklung der rechtlichen und infrastrukturellen Rahmenbedingungen. Weitere Empfehlungen des Gutachtens betreffen die Rahmenbedingungen für die wissenschaftliche Nutzung der Daten des FDZ Gesundheit allgemein sowie für die Linkage von Daten des FDZ Gesundheit mit weiteren VeDa wie den epidemiologischen und klinischen Krebsregisterdaten.

Das Gutachten stützt sich auf die langjährige Erfahrung der Gutachter in der wissenschaftlichen Nutzung und beim Linkage von GKV-Routinedaten mit anderen Datenquellen, sowie auf die breite Expertise des Gutachterteams in der Versorgungsforschung, Epidemiologie und evidenzbasierten Medizin. Im Rahmen der Gutachtenerstellung wurden für die Ableitung der Empfehlungen Gespräche mit den Ansprechpartnern des FDZ Gesundheit und der Vertrauensstelle am Robert-Koch Institut geführt. Die Perspektive von Patientenvertretenden, Forschenden, Krankenkassen und von Expertinnen und Experten für Medizinische Ethik wurde anhand von Fragebogenerhebungen (Patientenvertretende sowie Forschende) und von moderierten und aufgezeichneten Expertendiskussionsrunden (Ethik, Krankenkassen, Forschende) eingebunden. Die Protokolle und Zusammenfassungen der jeweiligen Erhebungen finden sich im Anhang.

1.2 Fallbeispiel: Das Nationale Pandemie Kohorten Netz (NAPKON) des Netzwerk Universitätsmedizin (NUM)

NUM -Netzwerk Universitätsmedizin

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Netzwerk Universitätsmedizin (NUM) hat zum Ziel, Daten, Erkenntnisse, Maßnahmenpläne, Diagnostik- und Behandlungsstrategien aller deutschen Universitätskliniken und ggf. weiterer Akteure im Sinne einer Pandemic Preparedness zusammenzuführen und auszuwerten (NUM, FKZ: 01KX2021). Es hat in einer ersten Förderphase dreizehn vordringliche Themen festgelegt, zu denen bundesweit klinikübergreifende Forschungsprojekte bearbeitet werden. Das Nationale Pandemie Kohorten Netz (NAPKON) ist das größte dieser Forschungsprojekte, es wird aktuell in einer zweiten Förderphase über weitere drei Jahren durch das BMBF gefördert.

Perspektivisch werden alle Universitätskliniken in Deutschland im NAPKON Netzwerk zusammenarbeiten. Sie werden die Daten der behandelten COVID-19-Patienten systematisch erfassen und in einer zentralen Datenbank bündeln. Zu dieser Datensammlung können zusätzliche Wissenschaftlerinnen und Wissenschaftler, die nicht in der medizinischen Forschung tätig sind, hinzugeholt werden. Sie werden mit ihrer Forschung die Arbeit der Ärztinnen und Ärzte flankieren. Koordiniert wird das Forschungsnetzwerk von der Charité – Universitätsmedizin Berlin. Das Bundesministerium für Bildung und Forschung (BMBF) fördert mit 150 Millionen Euro den Aufbau des Forschungsnetzwerks, um die Forschungsaktivitäten der deutschen Universitätsmedizin zur Bewältigung der COVID-19 Pandemie zu stärken. Der Bundestag hat Ende November 2020 mit der Annahme des Bundeshaushalts beschlossen, dass NUM auch nach 2021 in Höhe von 80 Millionen Euro jährlich, d. h. mit insgesamt 240 Millionen Euro bis zum Jahr 2024 weiter zu fördern. Im Erfolgsfall soll NUM anschließend verstetigt werden.

NAPKON

Das Nationale Pandemie Kohorten Netz (NAPKON) ist eine grundlegende Komponente des Netzwerks Universitätsmedizin und damit ein zentrales Projekt der Bundesregierung. Es wurde zu Beginn der COVID-19 Pandemie gestartet, um die Forschung und Versorgung durch die Unikliniken für Deutschland sicherzustellen. Im Rahmen von NAPKON werden erwachsene und pädiatrische SARS-Cov2 Patientinnen und Patienten in drei sich ergänzenden Kohortenplattformen (Sektorenübergreifende Plattform (SÜP), Hochauflösende Plattform (HAP) und Populationsbasierte Plattform (POP)) von der Erstinfektion bis zu einer Nachbeobachtungszeit von bis zu drei Jahren beobachtet. Zu den Verfahren der Primärdatenerhebung gehören eine umfassende klinische und bildgebende Diagnostik, die Beurteilung der Lebensqualität, von den Patienten berichtete Ergebnisse (patient-reported outcome measures, PROMS) und die Entnahme von Bioproben (Schons 2022). Die Datenerhebung erfolgt multizentrisch an den 36 Standorten der beteiligten Universitätskliniken und in der SÜP zusätzlich an außeruniversitären Krankenhäusern und bei niedergelassenen Ärztinnen und Ärzten. An den beteiligten Standorten wurde eine Beratung durch die Ethikkommissionen durchgeführt und positive Voten erteilt. Die in den drei NAPKON-Kohorten erhobenen Daten und Bioproben werden gemäß einer Nutzungsordnung auf Basis von Nutzungsanträgen über einen Use- und Access (UAC) Prozess geprüft und im positiven Fall den beantragenden Wissenschaftlerinnen und Wissenschaftlern für Forschungsvorhaben zur Verfügung gestellt.

Die drei NAPKON Primärdatenkohorten: POP, HAP und SÜP

Um den akuten Verlauf der COVID-19-Krankheit zu erfassen, aber auch um deren langfristigen Folgen zu analysieren, wurden im Rahmen des NUM-NAPKON-Projekts drei Kohorten etabliert, die verschiedene Abschnitte des Krankheitsverlaufs abbilden können: NAPKON-POP

(populationsbasierte Plattform), SÜP (sektorenübergreifende Plattform) und HAP (hochauflösende Plattform). Insgesamt steht den Forschern damit eine umfassende Sammlung von individuellen klinischen Daten, Bildgebung und Bioproben zur Verfügung [4].

Populationsbasierte Plattform (POP)

In der POP werden in geographisch definierten Sammelgebieten (Kiel, Berlin, Würzburg) alle Patientinnen und Patienten und Patienten mit überstandener SARS-CoV-2-Infektion schweregradunabhängig und populationsrepräsentativ identifiziert und in ein diagnostisches Langzeitprogramm eingeladen. In einem epidemiologisch ausgerichteten Fokuskrankenhaus der jeweiligen Sammelregionen erfolgt in drei lokal und strukturell unterschiedlichen deutschen Regionen die Rekrutierung [20]. Die Personen werden über die lokalen Gesundheitsbehörden identifiziert und kontaktiert, die verpflichtet sind, alle SARS-CoV-2-Infektionen in ihren Verwaltungsbezirken zu registrieren. Nach ihrer Zustimmung werden die in Frage kommenden Personen telefonisch befragt und anschließend zu einer Erstuntersuchung im Studienzentrum sowie zu jährlichen Nachuntersuchungen eingeladen. Die Besuche umfassen eine detaillierte retrospektive Erfassung des akuten Erkrankungsverlaufes und prospektive Biomaterialsammlung sowie die Erfassung der Langzeitfolgen und Komorbiditäten. Die POP ist beim Deutschen Register Klinischer Studien (DRKS) unter der Nummer DRKS00023742 registriert.

Sektorenübergreifende Plattform (SÜP)

Die SÜP rekrutiert SARS-CoV-2-infizierte stationäre und ambulante Patientinnen und Patienten und Patienten aller Altersgruppen über alle Abteilungen hinweg und führt eine umfassende Erhebung von primären Gesundheitsdaten, klinischen Basisphänotypisierungen (z. B. Echokardiographie, Spirometrie mit Ganzkörperplethysmographie) mit Bioprobenerhebung und ausführlichen Patientenbefragungen / patient-reported outcome measures (PROMs) auf allen Ebenen der Gesundheitseinrichtungen durch. Damit ist eine sektorenübergreifende Patientenerfassung im NAPKON gewährleistet. Alle deutschen universitären und außeruniversitären Krankenhäuser und Hausarztpraxen können Studienstandorte werden. Darüber hinaus sind mobile Hotspot-Studienteams für Langzeitpflege- und Rehabilitationseinrichtungen geplant. Die Kohorte ist bei www.clinicaltrials.gov unter NCT04768998 registriert [4].

Hochauflösende Plattform (HAP)

Die Kohorte der HAP konzentriert sich auf erwachsene SARS-CoV-2-positive stationäre Patientinnen und Patienten, insbesondere solche mit einem schweren Verlauf von COVID-19, d. h., die auf der Intensivstation behandelt werden müssen. Im Rahmen des HAP wird die Daten- und Bioprobensammlung durch ein multidisziplinäres Studienprogramm mit

zusätzlichen klinischen Untersuchungen, ergänzenden Zytokinprofilen und zusätzlicher standardisierter Bildgebung erweitert. Die longitudinale Bioprobenerhebung ist im Vergleich zu den beiden anderen Kohorten wesentlich häufiger. Die HAP-Kohorte wird an zehn deutschen Universitätskliniken durchgeführt, die über eine angemessene Infrastruktur für eine umfassende Phänotypisierung verfügen. Die Kohorte ist bei www.clinicaltrials.gov unter NCT04747366 registriert.

NUKLEUS (NUM Klinische Epidemiologie- und Studienplattform) umfasst die Infrastrukturen, die das Datenmanagement der NAPKON betreiben. NUKLEUS soll es ermöglichen, schneller und umfassender zu qualitätsgesicherter Forschung gelangen, um wichtige Erkenntnisse und Empfehlungen für die medizinische Versorgung der Bevölkerung zu erzeugen. Die gesamte Prozesskette der Studienunterstützung von Studiendesign, Einbindung der Studienzentren, Umsetzung der ethischen und rechtlichen Rahmenbedingungen, Erhebung des modularen Konsentes, Datentrennung und Pseudonymisierung, Qualitätssicherung, Studieninitiierung, Daten- und Bioproben-Management, faire Nutzung der gewonnenen Daten und Bioproben, Datenanalyse und Datennachnutzung ist in NUKLEUS transparent geregelt [21]. NUKLEUS besteht aus den 3 Methodenkernen Interaktionskern (ICU), Epidemiologiekern (ECU), Bioprobenkern (BCU) und den Infrastrukturen Transferstelle (TO), Klinisches Datenmanagement (CDM), Bilddatenmanagement (DIMA), Bioproben-Information Management System (BIMS), Ethik-Koordination (EC) und der Treuhandstelle für Konsent-, ID und Pseudonymmanagement (THS).

NAPKON Use and Access Prozess

Das NAPKON Use & Access Committee (UAC) wurde Ende 2020 gegründet. Stimmberechtigte Mitglieder des UAC sind acht aus dem Fachbeirat durch die Vollversammlung gewählte Personen, je eine Repräsentantin bzw. einen Repräsentanten der drei NAPKON Kohortenplattformen, des Bioprobenkerns sowie der Epidemiology Core Unit. Die beratenden Mitglieder sind je eine Vertreterin bzw. ein Vertreter der Ethik- und der Patientenperspektive und werden vom NAPKON-Lenkausschuss ernannt. Seit der Eröffnung des Use & Access-Prozesses am 26. April 2021 hat der UAC die eingegangenen Projektvorschläge in wöchentlichen Sitzungen diskutiert und darüber abgestimmt. In den einzureichenden NAPKON UAC Antragsunterlagen wird explizit die Frage nach einer Datenverknüpfung gestellt, um mögliche Implikationen im UAC Komitee besprechen zu können. Ein positives Ethikvotum ist für jedes Projekt vorhanden.

NUM FOSA

Die Fach- und Organspezifischen Arbeitsgruppen (FOSA) dienen dem inner- und interdisziplinären Austausch der verschiedenen medizinischen Fach- und Organrichtungen im NUM und haben wichtige Funktionen innerhalb des Netzwerks: Sie kommentieren Datensätze und unterziehen sie einer Qualitätssicherung und ggfs. Revision, sie nutzen Daten- und Bioproben für die Forschung, sie kommentieren Anträge und beteiligen sich am Use & Access

Verfahren. Außerdem bilden die FOSA-Sprecherinnen und –sprecher den Fachbeirat, der regelmäßig tagt und auf interdisziplinärer Ebene die Weiterentwicklung von NUM unterstützt. Die FOSA stehen Ärztinnen und Ärzten und Wissenschaftlerinnen und Wissenschaftlern offen. Die ca. 1.500 in den FOSA organisierten Wissenschaftlerinnen und Wissenschaftlern wurden stellvertretend für weitere NUM Projekte über deren Fachbeirat für das vorliegende Gutachten zu ihren Meinungen bzgl. Datenverknüpfung, Forschungsbedarfen sowie zu Barrieren und Fördernden Faktoren für die Nutzung von GKV-Daten im FDZ und zum Datenlinkage befragt (Anlage 2) [21].

1.3 Machbarkeitsstudie im Fallbeispiel Datenverknüpfung der klinischen NAPKON-POP Daten mit GKV Routinedaten

Die Verknüpfung individueller klinischer Primärdaten aus den drei NAPKON-Kohorten mit GKV-Routinedaten auf der Basis einer informierten Einwilligung ermöglicht eine viel umfassendere Analyse von Krankheits- und Versorgungsprozessen bei COVID-19-Erkrankungen als dies mit klinischen Daten allein möglich wäre. Daher wurde die Möglichkeit einer solchen Datenverknüpfung bereits zu Beginn der NAPKON-Studien in der modularisierten Einverständniserklärung und Patienteninformation verankert. Insbesondere in der bevölkerungsbezogenen Kohorte (NAPKON-POP) wurden die Krankenversicherungsnummer und die Krankenkassennummer von Anfang an im Studienzentrum Würzburg erfasst (für das vollständige Studienprotokoll siehe Anlage 1a).

Ziele und Fragestellungen der Fallstudie mit dem Titel: „Record-Linkage von Individualdaten einer epidemiologischen Kohortenstudie (NAPKON-POP - COVIDOM Kohorte) mit GKV-Routinedatendaten: Eine Proof-of-Concept-Studie mit SARS-CoV-2 infizierten Personen nach der akuten Infektion.“

GKV-Routinedaten ermöglichen die vollständige und sektorübergreifende Abbildung von Krankheitsverläufen einschließlich Informationen über Vorerkrankungen sowie langfristige Nachbeobachtungszeiträume (bis zu mehreren Jahren) und Analysen der gesundheitsökonomischen Kosteneffizienz der Versorgung. Auf der Grundlage hochwertiger Evidenz aus der Versorgungsforschung können anschließend wirksame Versorgungskonzepte für die COVID-19-Krankheit entwickelt, umgesetzt und evaluiert werden.

Ziel der ersten Phase ist es, die Machbarkeit der auf einer informierten Einwilligung (informed consent-) basierten Datenverknüpfung für die Teilnehmende der NAPKON-POP Kohorte zu belegen, harmonisierte Prozesse des gesamten Datenflusses zu etablieren und eine erste ergänzende, orientierende Validierung der verknüpften Daten vorzunehmen.

Die erste Phase (Machbarkeitsstudie) kann aufgrund unterschiedlicher Barrieren, welche auch einen wesentlichen Teil dieses Gutachtens darstellen, zurzeit nicht weitergeführt werden.

Zentral ist, dass derzeit die dazu notwendige Bereitschaft von einzelnen Krankenkassen fehlt, das Projekt durch eine Datenlieferung für das Datenlinkage zu unterstützen. Eine erneute Ansprache der Krankenkassen wird erfolgen, wenn eine größere Anzahl von Patienten mit vorliegender Einwilligung, einschließlich der KVNR, in die NAPKON Studie rekrutiert worden ist.

Die Hautfragestellungen der Machbarkeitsphase sind:

1. Ist eine Datenverknüpfung von GKV-Routinedaten kooperierender Krankenkassen mit den NAPKON-Pop-Kohortendaten möglich?
2. Welche Ergänzungen der bestehenden NUM-Infrastruktur sind notwendig, um die Voraussetzungen für ein datenschutzkonformes und praktikabel einsetzbares Datenlinkage von GKV-Daten und primären Kohortendaten zu schaffen?
3. Welche prozessbezogenen Hindernisse bestehen bzgl. des vorgesehenen Datenlinkage?
4. Wie groß ist der personelle, zeitliche und IT-technische Aufwand für die (routinemäßige) Umsetzung?
5. In welchen inhaltlichen Bereichen und zu welchem Grad können die erhobenen Daten zum Krankheitsverlauf und zum Versorgungsgeschehen beider Datenquellen wechselseitig validiert werden?
6. Welche Variablen und welche operationalen Schritte sind zur Anwendung bestehender Definitionen von "Long-Covid" nötig und welche davon sind in den zusammengeführten Daten möglich?

Nach erfolgreicher Umsetzung und Auswertung der Proof-of-Concept-Phase in der NAPKON-POP Kohorte werden die entwickelten Prozesse auf alle drei NAPKON-Kohorten ausgerollt.

Die gewonnenen Erkenntnisse werden im Rollout (Phase 2) umgesetzt, um umfassende Daten zur Gesundheitsversorgung der Teilnehmende aller NAPKON-Kohorten durch Verknüpfung der NAPKON-Untersuchungsdaten mit GKV-Routinedaten mit vor und nach ihrer der SARS-CoV-2-Infektion sowie vor und nach der Pandemie für die Forschung bereit zu stellen. Eine realistische Umsetzung des Linkage unter Einbezug der Daten aller gesetzlicher Krankenkassen erfordert die Funktionalität des FDZ.

Methodisch interessant ist auch die Ziehung einer Kontrollgruppe aus nicht-identifizierten Personen ohne SARS-CoV-2-Infektion ausschließlich auf Basis der GKV-Routinedaten, beispielweise durch Propensity-Score-Matching. Diese Vorgehensweise ermöglicht bei entsprechendem Design der Kontrollgruppe Unterschiede hinsichtlich in der Inanspruchnahme (Inanspruchnahme bestimmter Facharztgruppen) oder Unterschiede in der

Komorbidität/Komedikation darzustellen. Diese vergleichenden Analysen benötigen im Datensatz des FDZ Gesundheit notwendigerweise die Markierung der KohortenTeilnehmende, um aus dem Pool der NichtTeilnehmende die Kontrollgruppe definieren zu können.

Die Rollout-Phase wird begleitet durch ein engmaschiges Berichtssystem der einzelnen NAPKON Kohorten und intensiven Erfahrungsaustausch zwischen den Kohorten. Ziel ist, möglichst frühzeitig notwendige Schnittstellenprobleme zu erkennen und Vorschläge für ein harmonisiertes Vorgehen zu entwickeln.

Projektfortschritt der Fallstudie und Erfahrung bei der Umsetzung:

Im NAPKON Projekt wurde eine Arbeitsgruppe ‚GKV Daten‘ eingerichtet, um die Umsetzung der Datenverknüpfung zu unterstützen. Jedes Verknüpfungsprojekt ist auf die Zusammenarbeit und Abstimmung mit den gesetzlichen Krankenkassen angewiesen, bei denen die Studienteilnehmenden versichert sind. Bei Vorliegen der sozial- und datenschutzrechtlichen Voraussetzungen ist es die Entscheidung der Krankenkassen, ob sie ein Projekt unterstützen und welche Daten sie zur Verfügung stellen. Die Daten des zeitlichen Ablaufs der Fallstudie zum aktuellen Stand sind in Kapitel 2.4. tabellarisch abgebildet.

Zunächst wurde mit ausgewählten Krankenkassen Kontakt aufgenommen, um die Inhalte der NAPKON-übergreifenden Einverständniserklärungen abzustimmen. Die Abstimmungsgespräche führten zur Einführung einer angepassten und von den ursprünglichen Dokumenten differenzierteren Version der Einverständniserklärungen/Patienteninformation im gesamten NAPKON (Anlage 1b).

Insbesondere die Forderung der Krankenkassen, die wissenschaftlichen Fragestellungen konkreter zu formulieren und den Zeitrahmen weiter zu spezifizieren, für die Daten analysiert werden sollen, sollten in den Amendments umgesetzt werden.

Der Projektplan für die Modelllösung für NAPKON sieht den Standort NAPKON-POP in Würzburg als Ausgangspunkt vor, da hier neben den für die Verknüpfung der Daten eingeholten Einverständniserklärungen bereits die Krankenkassennummern (KVNR) als Voraussetzung der Versichertenidentifikation durch die GKVn erhoben wurden. Auch die Krankenversicherungsnummer wurde vermerkt. Andere NAPKON-Kohorten und -Standorte ziehen mit diesen Schritten nach.

Die Beurteilung der Einwilligungunterlagen durch die zuständigen Ethikkommissionen war dabei durchaus heterogen: Während die Amendments, welche ja genau zum Zwecke einer Spezifizierung gegenüber früheren Versionen implementiert wurden, von einigen EKs nicht beanstandet wurden, lehnte eine EK eine Erfassung von Gesundheits- und Sozialdaten in der HAP Kohorte mit folgender Begründung ab:

„Es sollen Gesundheits- und Sozialdaten erfasst werden. Lediglich aus den Studieninformation ergibt sich, dass der Zeitraum bis zu 5 Jahre nach Infektion sein soll. Dies findet keine Zustimmung. Eine Datenerfassung hoch sensibler Daten für einen Zeitraum der im Zeitpunkt der Einwilligung so weitreichend wäre und in einer Breite, die für die Beantwortung der Forschungsfrage nicht erforderlich ist, ist nicht einwilligungsfähig. Die Ethikkommission kann der Erhebung dieser Daten nur zustimmen sofern diese bereits erhoben sind, da dies für die Teilnehmenden ein überschaubarer Zeitraum ist, bei dem sie auch wissen, welche Daten sie preisgeben. Es wird empfohlen eine eigene Studie mit eingeschränkten Datensätzen zu den [...] genannten Fragestellungen zu planen und einzureichen.“

Auch nach einer erläuternden Erwiderung der Studienkoordination blieb die Ethikkommission mit folgendem Passus bei ihren ursprünglichen Bedenken und begründet weiter:

„Ein spezifisches Widerspruchsrecht ist nicht ausreichend, da der typische Umfang bei der Nutzung von Gesundheitsdaten in Studien den wenigsten Patientinnen und Patienten bekannt sein dürfte und sie die Konsequenzen nur schwer abschätzen können, gerade da die Teilnehmenden die zukünftigen Datenabfragen auch nicht (mehr) bewusst mitbekommen.“

Vor dem Hintergrund heterogener Beurteilungen durch die Ethikkommission konnten die entsprechenden Versionen der Einwilligungsunterlagen nicht an allen Standorten der NAPKON Kohorten umgesetzt werden.

Zum weiteren Projektplan für die Ausarbeitung einer Modelllösung gehörte auch die Planung einer Proof-of-Concept-Studie, für die ein detailliertes Studienprotokoll entwickelt wurde. Dieses beinhaltet ein Datenschutzkonzept unter Einbeziehung einer Treuhandstellenlösung, die detaillierte Beschreibung der zu realisierenden Datenflüsse mit einem Datenannahme- und Datenverarbeitungszentrum sowie die Entwicklung einer GKV-Routinedatenbank unter Berücksichtigung der bereits vorhandenen NUM-Strukturen. Das Studienprotokoll findet sich im Anhang (Anlage 1a); es enthält auch eine ausführliche Definition des GKV Datensatzes (minimal data set), der zur Beantwortung der Forschungsfragen (siehe Kapitel 1.2) und zur Untersuchung der Machbarkeit benötigt wurden.

Als schwierig erwies sich auch die Kontaktaufnahme mit den Krankenkassen für eine Kooperation. Angestrebt war, dass die Krankenkassen GKV-Routinedaten ihrer Versicherten wie in der Patienteninformation beschrieben auf der Basis einer informierten Einwilligung zur Verfügung stellen. Wie in solchen Kooperationsprojekten üblich wurden die Kassen angefragt, sich mitgestaltend bei der Durchführung des Projekts zu beteiligen, insbesondere auch bei Rückfragen zu Dateninhalten oder im Hinblick auf die Interpretierbarkeit der Ergebnisse aus Kassensicht. Gleichzeitig wurde den kontaktierten Krankenkassen auch die Möglichkeit eröffnet, im Rahmen der Kooperation über das NAPKON Use and Access Verfahren selbst Anträge für die wissenschaftliche Nutzung der gelinkten Daten einzubringen. Es war im

NAPKON Projekt zunächst unklar, wie erfolgreiche Kooperationsanfragen an verschiedene Krankenkassen gestellt werden sollten. Aus diesem Grund mussten externe Experten (u. a. aus der Expertengruppe) als Berater und zur Kontaktaufnahme hinzugezogen werden. Auf der Grundlage früherer gemeinsamer Forschungsprojekte und persönlicher Kontakte wurden verschiedene Krankenkassen mit besonders vielen Patienten in der Population der Machbarkeitsstudie angesprochen und um eine Beteiligung am Projekt gebeten. Nach dem derzeitigen Stand ist es trotz mehrerer direkter Anfragen und erläuternder Gesprächsrunden nicht gelungen, eine Krankenkasse für eine Zusammenarbeit zu gewinnen. Die Erfahrungen aus der Vorbereitung der Proof-of-Concept Studie finden sich zusammengefasst in Kapitel 2.5.

Die geplante Datenbasis für die Datenverknüpfung und die anschließende Analyse bilden (1) Identifikationsdaten (ID) der Kohorten von NAPKON, die über die NUKLEUS-Strukturen der Datenhaltung, der Unabhängigen Treuhandstelle und der Datentransferstelle verwaltet werden, sowie (2) Identifikationsdaten (ID) der ambulanten und stationären Verordnungs- und Abrechnungsdaten ausgewählter Krankenkassen. Die Identifikationsdaten werden ausschließlich an die Treuhandstelle übermittelt und nach Zusammenführung mit den ID-Daten der Kohortenteilnehmenden pseudonymisiert. Nur die pseudonymisierten Datensätze werden in die Datenhaltung übernommen und stehen für die Analyse zur Verfügung. In die NAPKON-POP [20] wurden bis 31. 12. 2021 über 2000 Patientinnen und Patienten an den drei Standorten rekrutiert (Kiel: 1663, Berlin: 302, Würzburg 348). In der laufenden zweiten Förderperiode (2022-2024) sollen weitere ca. 1.200 Patientinnen und Patienten an den drei Standorten rekrutiert werden (Kiel: ca. 700, Berlin: ca. 250, Würzburg ca. 250).

Das Studienprotokoll (Anlage 1a) beschreibt detailliert die Elemente der NAPKON-POP-Basisuntersuchung und der Nachuntersuchung nach 12 Monaten für ausgewählte Patienten mit persistierenden COVID-19-Symptomen sowie die Ein- und Ausschlusskriterien. Eine detaillierte Beschreibung aller NAPKON-Kohorten findet sich in [4].

Die zweite Datenquelle zur Verknüpfung besteht aus pseudonymisierten ambulanten und stationären Verordnungs- und Abrechnungsdaten ausgewählter Krankenkassen, deren detaillierte Variablenbeschreibung ebenfalls im Studienprotokoll zu finden ist. Nach erfolgreicher Testung der Prozesse in der Machbarkeitsphase sollen nach initialer Einwilligung der Patienten deren entsprechende Verordnungs- und Abrechnungsdaten von deren Krankenkasse eingeholt und in der Datenhaltung gespeichert werden. Die Datenverknüpfung erfolgt unter Einbindung einer Treuhandstelle (THS), die den Identitätsabgleich zwischen den verschiedenen Datenquellen vornimmt, aber keine zusätzlichen Daten im Sinne des Trennungsgebots erhält und die identifizierenden Informationen nicht weitergibt. Lediglich eine Zuordnung von Transportpseudonymen wird an die Datenannahme- und Aufbereitungsstelle weitergegeben. Um den Datenschutz zu gewährleisten, sind an der Datenannahme- und Aufbereitungsstelle am ZEGV Dresden und an der TU Dresden getrennte Stellen für die Datenverwaltung und Datenauswertung ohne gegenseitige Weisungsbefugnis

eingrichtet worden. Die beteiligten Projektmitarbeitenden des ZEGV arbeiten auf der Grundlage eines einrichtungsspezifischen Datenschutzkonzepts, das auf die Einhaltung der datenschutzrechtlichen Bestimmungen, sowie der guten wissenschaftlichen Praxis und der guten Praxis Sekundärdatenanalyse [22, 23] abzielt. Die Einhaltung der Standard Operating Procedure (SOP)s Datenschutz und Datensicherheit sind für alle Projektmitarbeitende obligatorisch. Die SOP enthält u. a. Angaben zur personellen Verantwortlichkeit, zum Zugriffsschutz, zur Datensicherung und zur Verfügbarkeitskontrolle. Die Sicherung und der Schutz von Sozialdaten sind damit nach den Anforderungen gemäß § 78a SGB X geregelt.

Die pseudonymisierten Daten der Krankenkassen werden von der Datenhaltung entgegengenommen, zusammengeführt, harmonisiert und auf Plausibilität und Konsistenz geprüft. Werden die Daten auf Antrag für eine spezifische Fragestellungen angefordert, so bereitet die Datenaufbereitungsstelle die Daten entsprechend der projektspezifischen Anforderungen auf. In der Regel wird hierbei eine Zielpopulation über Ein-/Ausschlusskriterien definiert, für die ein projektspezifischer Variablensatz bereitgestellt wird, der nur die für die Beantwortung der Forschungsfrage notwendigen Tabellen, Zeilen und Spalten sowie Ausprägungen enthält. Die Aufbereitungsstelle leitet die projektspezifischen Daten damit lediglich als Auszug aus den Versorgungsdaten der Krankenkassen weiter. Als Personenmerkmal wird eine projektspezifische TransPID1 als Studienidentifikator verwendet und mitübermittelt (siehe Abbildung 1). Dieser Identifikator dient auch als Schlüsselvariable für die Verlinkung der beiden Datenauszüge aus der Nukleus-Transferstelle (klinischen) Daten und der Datenannahme und Aufbereitungsstelle ZEGV, die beide dieselbe TransPID1 als Studienidentifikator nutzen. Eine Rückübermittlung der TransPID1 ist nicht vorgesehen und eine Reidentifizierung damit nicht möglich. Für den Datennutzer sind diese nicht reidentifizierbaren Daten damit „faktisch anonymisiert“.

Die ID-Daten und pseudonymisierten medizinischen Daten der NAPKON-Kohorte werden gemäß den Studienprotokollen erhoben und verarbeitet, die von den zuständigen Ethikkommissionen begutachtet wurden. Die Verwendung und Zusammenführung der ID-Daten in der Treuhandstelle ist durch die Patienteninformationen und Einverständniserklärungen (Informed Consent) der NAPKON-Kohorten abgedeckt. Die verwendeten modularen Einwilligungserklärungen finden sich in Anlage 1c. Für die NAPKON-POP Standorte liegen derzeit modulare Einwilligungen zur Datenzusammenführung, Kontakt- und Versicherungsdaten für ca. 600 Probandinnen und Probanden sowie ein positives Ethikvotum der Medizinischen Fakultät der Universität Würzburg vor (236/20_z).

Die Datennutzung erfolgt gemäß der NAPKON Nutzungsordnung und den darin festgelegten Use & Access Regularien [24]. Entsprechend wird die Datennutzung beim NAPKON Use und Access Komitee (UAC) beantragt.

Die Datenherausgabe erfolgt dann über eine Datentransferstelle (für Details siehe Diagramm im Studienprotokoll).

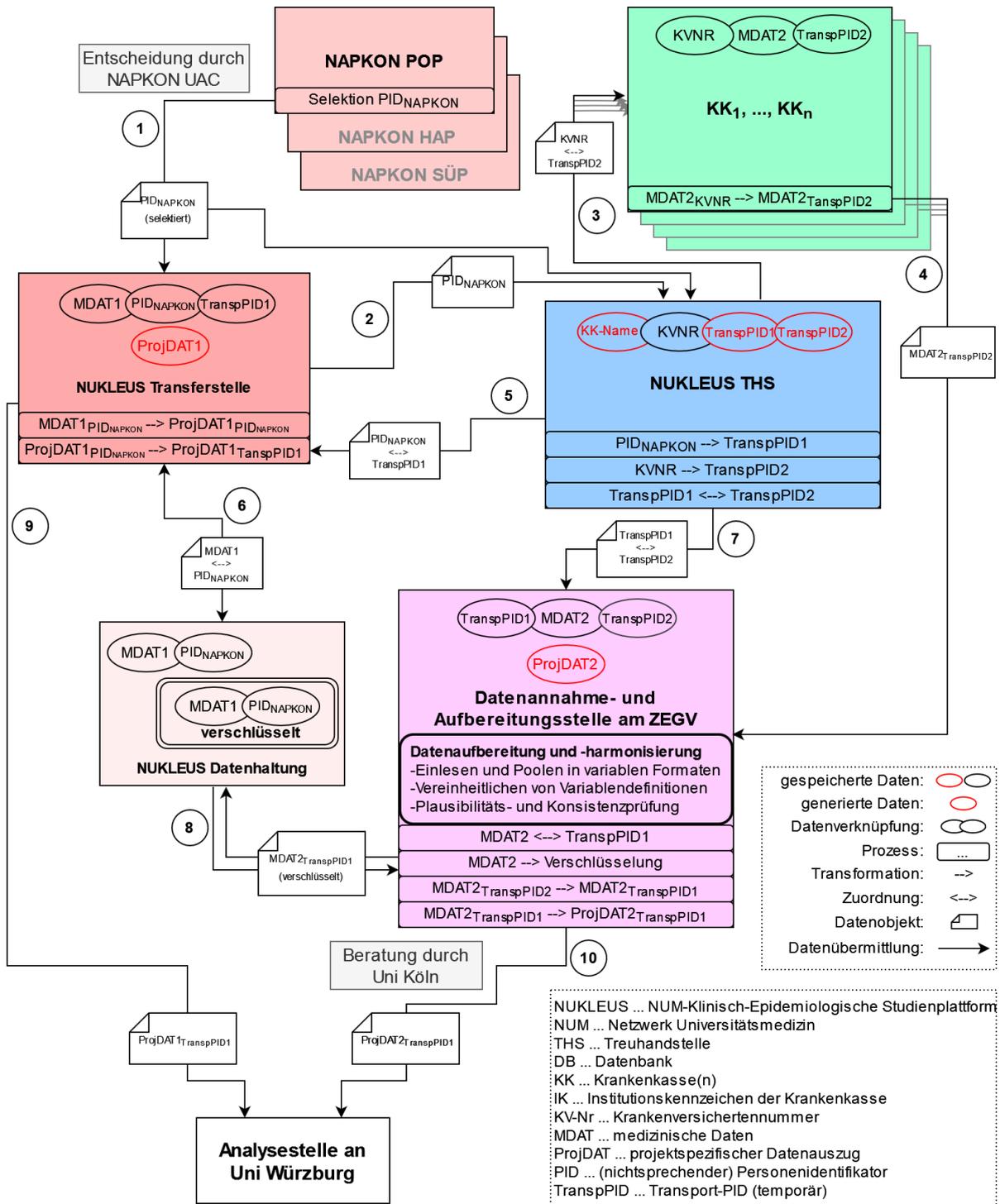


Abbildung 1: Datenfluss, Datenverknüpfung und Datenanalyse - Proof-of-Concept unter Beteiligung spezifischer Krankenkassen

Tabelle 2: Legende zu den nummerierten Datenflüssen aus Abbildung 1

Nummer	verantwortliche Stelle	Datenübermittlung
(1)	NAPKON POP	Übermittlung des NAPKON-Pseudonyms von NAPKON-Studienteilnehmenden mit Einwilligung zur Verknüpfung der Studiendaten mit Krankenkassendaten, an die NUKLEUS Treuhandstelle, um diese Studienteilnehmenden für nachfolgende Datenflüsse selektieren zu können.
(2)	NUKLEUS Treuhandstelle	Prüfung des aktuellen Einwilligungsstatus der NAPKON Kohortenteilnehmenden in das Linkage mit GKV-Daten. Die NUKLEUS Treuhandstelle vergibt das projektspezifische NAPKON-Pseudonym für die Anfrage bei den kooperierenden Krankenkassen.
(3)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt aus der KVNR ein Transport-Pseudonym (TranspPID2) und stellt dieses zusammen mit einer Zuordnung der vorliegenden KVNR den verschiedenen Krankenkassen zur Verfügung. Dabei erhält jede Krankenkasse nur die KVNR-TranspPID2-Kombinationen für NAPKON-Studienteilnehmende, welche in der jeweiligen Krankenkasse versichert sind.
(4)	Krankenkassen	Die Krankenkassen selektieren die ihnen vorliegenden medizinischen Daten (MDAT2) auf die übermittelten KVNR sowie den vorab vereinbarten Datenkranz, ersetzen für diesen Datenauszug die KVNR durch das Transport-Pseudonym (TranspPID2) und stellen den so bearbeiteten Datenbestand (MDAT2 _{TranspPID2}) der Datenannahme- und Aufbereitungsstelle am ZEGV zur Verfügung.
(5)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt aus dem NAPKON-Pseudonym von NAPKON-Studienteilnehmende mit Einwilligung ein projektspezifisches Pseudonym (TranspPID1) und stellt dieses zusammen mit einer Zuordnung des NAPKON-Pseudonym der NUKLEUS Transferstelle zur Verfügung.

Nummer	verantwortliche Stelle	Datenübermittlung
(6)	NUKLEUS Transferstelle und Datenhaltung	Die medizinischen NAPKON-Daten (MDAT1) werden zusammen mit dem NAPKON-Pseudonym in der NUKLEUS Datenhaltung gespeichert, anlassbezogen in der NUKLEUS Transferstelle verarbeitet und zwischen beiden Stellen übermittelt.
(7)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt für NAPKON-Studienteilnehmende mit Einwilligung eine Zuordnungstabelle zwischen dem projektspezifischen Pseudonym (TranspPID1) sowie dem Transport-Pseudonym (TranspPID2) und stellt diese der Datenannahme- und Aufbereitungsstelle am ZEGV zur Verfügung.
(8)	Datenannahme- und Aufbereitungsstelle am ZEGV und NUKLEUS Datenhaltung	Die Datenannahme- und Aufbereitungsstelle am ZEGV poolt, harmonisiert sowie plausibilisiert die von den Krankenkassen erhaltenen Daten und ersetzt das Transport-Pseudonym durch das projektspezifische Pseudonym. Eine verschlüsselte Version der so aufbereiteten Daten übermittelt die Datenannahme- und Aufbereitungsstelle an die NUKLEUS-Datenhaltung, jedoch ohne den Schlüssel zur Entschlüsselung dieser Daten. Nach Schritt (10) werden die unverschlüsselten Daten in der Datenannahme- und Aufbereitungsstelle gelöscht. Für eine gegebenenfalls erforderliche Korrektur der Aufbereitung werden die verschlüsselten Daten aus der NUKLEUS Datenhaltung an die Datenannahme- und Aufbereitungsstelle übermittelt und dort mit dem vorliegenden Schlüssel entschlüsselt.

Nummer	verantwortliche Stelle	Datenübermittlung
(9)	NUKLEUS Transferstelle	Nachdem die NAPKON-Daten in der NUKLEUS Transferstelle auf den projektspezifischen Datensatz von NAPKON-Studienteilnehmende mit Einwilligung sowie benötigten Analyse-Variablen eingeschränkt wurden (ProjDAT1), wird das NAPKON-Pseudonym durch das projektspezifische Pseudonym (TanspPID1) ersetzt und dieser Datensatz (ProjDAT1 _{TanspPID1}) von der NUKLEUS Transferstelle an die Analysestelle übergeben.
(10)	Datenannahme- und Aufbereitungsstelle am ZEGV	Die Datenannahme- und Aufbereitungsstelle am ZEGV erstellt aus den MDAT2 einen projektspezifischen Datenauszug (ProjDAT2 _{TanspPID1}) und stellt diesen an die Auswertestelle an der Uni Würzburg zur Verfügung. Die Auswertestelle kann die beiden projektspezifischen Datensätze ProjDAT1 _{TanspPID1} und ProjDAT2 _{TanspPID1} anhand des projektspezifischen Pseudonyms TranspPID1 verknüpfen und gemeinsam analysieren.

An dieser Stelle werden Datenfluss, Datenverknüpfung und Datenanalyse schematisch abgebildet. Eine detaillierte Beschreibung findet sich im Studienprotokoll im Anhang (Anlage 1a). Üblicherweise können die notwendigen Sekundärdaten von den angefragten Krankenkassen mit einer Latenz von 1 bis 4 Monaten für die stationären und von 7 bis 9 Monaten für die ambulanten Daten bereitgestellt werden. Bei dem zunächst vorgesehenen Zeitplan für die Datenverknüpfung und verfügbarer Daten im 4. Quartal 2022 wären somit umfassende Analysen für einen Zeitraum vor Pandemiebeginn (ab Quartal 1, 2019) für das gesamte Jahr 2021 möglich gewesen.

Der zeitliche Verlauf der zu implementierenden Schritte und Prozesse wird im Folgenden beschreiben:

Zu Beginn erfolgt der Einschluss von Patientinnen und Patienten in der NAPKON-POP Kohorte mit Einwilligung zur Erfassung der Krankenkassendaten: Sobald eine Patientin bzw. ein Patient in der NAPKON-POP-Kohorte in die Verknüpfung mit den Krankenkassendaten einwilligt, wird bei der NUKLEUS Treuhandstelle zusätzlich die KVNR der Patientin/des Patienten und ein Kennzeichen der jeweils versichernden KK erfasst (z. B. Name der Kasse).

Der Prozess der Beantragung und der Herausgabe der Krankenkassendaten einer spezifizierten Kohorte von Patientinnen und Patienten beginnt mit der Antragsstellung von Forschenden über das NAPKON Use&Access Komitee zur Herausgabe von Krankenkassendaten zu einer spezifizierten Kohorte bzw. einer Subgruppe selektierter Teilnehmende. Nach positivem Votum des UAC und Vorliegen eines projektspezifischen Ethikvotums wird über die NUKLEUS Transferstelle für die definierten Befragungsteilnehmenden/Patientinnen und Patienten in der Treuhandstelle der Einwilligungsstatus zur Datenverknüpfung über das das zugehörige NAPKON-Pseudonym (PID_{NAPKON}) geprüft.

Bei aktuell vorliegender Einwilligungserklärung in die Datenverknüpfung erzeugt die Treuhandstelle aus dem NAPKON-Pseudonym zwei unterschiedliche systemfreie personenbezogene Transport-IDs (TranspPID1, TranspPID2), ordnet die in der Treuhandstelle vorliegende KVNR der TranspPID2 zu und übermittelt die KVNR und die TranspPID2 inklusive Zuordnungsvorschrift an die jeweils versichernde KK. Zudem übermittelt die Treuhandstelle die TranspPID1 und die TranspPID2 inklusive Zuordnungsvorschrift an die Datenannahme- und Aufbereitungsstelle am ZEGV Dresden sowie die PID_{NAPKON} und TranspPID1 inklusive Zuordnungsvorschrift an die NUKLEUS Datenhaltung. Die THS übernimmt die Ansprache der KKen und steht für Rückfragen von den KKen zur Verfügung Die Übermittlungswege zu den Krankenkassen, sowie die zu übermittelnden Daten an die Krankenkassen werden kassenspezifisch mit jeder Einrichtung auf Basis ihrer datenschutzrechtlichen und technischen Vorgaben vereinbart. Die Krankenkassen selektieren die Daten (MDAT2) der bei ihnen versicherten Befragungsteilnehmende mit Einwilligung zur Datenverknüpfung anhand der KVNR, ersetzen die KVNR durch die TranspPID2 und übermitteln diesen pseudonymisierten Datensatz an die Datenannahme- und Aufbereitungsstelle am ZEGV. Datenlieferungen und Verknüpfungen werden mit der Einwilligung eines NAPKON-Teilnehmers initiiert. In dieser Einwilligung ist das Spektrum möglicher Auswertungen beschrieben und kann ggfs. von den NAPKON-Teilnehmenden spezifisch eingeschränkt werden. Von den Krankenkassen wird in der Folge jeweils ein Standardsatz von Variablen angefordert, der möglichst alle Auswertungen innerhalb des eingewilligten Spektrums ermöglicht. Ziel ist, in nachfolgenden Use & Access-Verfahren möglichst keine projektspezifischen Variablen von den Krankenkassen anzufordern. Der übergebene Standardvariablensatz wird somit einzelprojektprojektübergreifend gespeichert. Analog würde bei Datenanfragen und -übergaben auf der Basis des Broad Consent der Medizininformatik-Initiative verfahren.

Die Datenanforderung wiederholt sich in der Regel jährlich über den jeweiligen Einwilligungszeitraum. Nach Plausibilitätsprüfung und Aufbereitung werden die GKV-Daten an die NUKLEUS Datenhaltung weitergegeben und dort analog zu den Kohortendaten vorgehalten, mit dem Ziel, dass die GKV-Daten immer abrufbereit zur Verfügung stehen, so dass sie zusammen mit den jeweils benötigten Kohortendaten für begutachtete und genehmigte spezifische Auswertungsprojekte über die NUKLEUS-Transferstelle herausgegeben werden können.

Die Datenannahme- und Aufbereitungsstelle (ZEGV, Hochschulmedizin Dresden) führt die Daten der verschiedenen Krankenkassen zusammen und ersetzt anhand der vorliegenden Zuordnungsvorschrift die TranspPID2 durch die TranspPID1. Zudem werden die aus den verschiedenen Krankenkassen mit unterschiedlichem Datenformat bereitgestellten Daten hinsichtlich Datenformat und Variablendefinitionen harmonisiert, auf Plausibilität und Konsistenz überprüft und die für das Forschungsvorhaben benötigten Variablen zur Weitergabe selektiert (ProjDAT2). Um keine dauerhafte Datenhaltung der MDAT2 am ZEGV zu etablieren, aber auf Rückfragen und Nachforderungen ohne erneute Datenlieferung der unterschiedlichen KK angemessen reagieren zu können, werden nach Extraktion der ProjDAT2 die aufbereiteten und harmonisierten MDAT2 mit zugeordneter TranspPID1 verschlüsselt an die NUKLEUS Datenhaltung übergeben und im ZEGV anschließend gelöscht. Der Schlüssel zum Entschlüsseln der Daten verbleibt im ZEGV. So hat keine beteiligte Stelle dauerhaften Zugriff auf die aufbereiteten MDAT2, können aber anlassbezogen die verschlüsselten Daten an das ZEGV rückübermittelt, dort entschlüsselt und für weitere Datenaufbereitungen genutzt werden.

Die Datenanalyse wird (im Fallbeispiel der Machbarkeitsstudie) anhand der verknüpften Daten nach Entscheidung des NAPKON UAC bezüglich der Nutzung von NAPKON POP Daten und nach Beratung durch die Universität Köln bezüglich der Nutzung von Krankenkassen Daten an der Analysestelle an der Universität Würzburg durchgeführt. Die NAPKON NUKLEUS Datenhaltung sowie die Datenannahme- und Aufbereitungsstelle erstellen jeweils einen auf das Forschungsprojekt zugeschnittenen Auszug der jeweils vorliegenden NAPKON-POP- bzw. KK-Daten (MDAT1 und MDAT2, jeweils pseudonymisiert mit Bezug zu TranspPID1) und übermitteln diese an die Analysestelle. Diese kann anhand der gleichartigen TranspPID1 die beiden Datenquellen verknüpfen und gemäß der Use & Access-Entscheidung und Studienprotokoll verknüpft auswerten. Alle beschriebenen Datenflüsse zwischen verschiedenen beteiligten Stellen erfolgen geschützt (verschlüsselt und/oder über speziell geschützte Transportwege).

Das Projekt erfolgt im Einklang mit der Guten Epidemiologischen Praxis (GEP), der Guten Praxis Datenlinkage (GPD), den Empfehlungen des Datenschutzeitfadens der TMF für medizinische Verbundforschungsprojekte [22, 23]. Die Berichterstattung erfolgt unabhängig vom Projektergebnis nach den aktuellen Berichtsstandards für Sekundärdatenanalysen.

2 Linkage von Forschungsdaten mit GKV-Daten

2.1 Forschungsbedarfe / Fragestellungen

Nach § 12 SGB V müssen Versorgungsleistungen "(...) ausreichend, zweckmäßig und wirtschaftlich sein; sie dürfen das Maß des Notwendigen nicht überschreiten. Leistungen, die nicht notwendig oder unwirtschaftlich sind, können Versicherte nicht beanspruchen, dürfen

die Leistungserbringer nicht bewirken und die Krankenkassen nicht bewilligen." Hieraus resultieren für die Untersuchung versorgungsrelevanter Fragestellungen komplexe Anforderungen. Nicht nur - wie in klinischen Studien - die Wirksamkeit oder - wie in gesundheitsökonomischen Studien - die Effizienz müssen in den Analysen berücksichtigt werden, sondern darüber hinaus der Bedarf (Notwendigkeit) und Aspekte der Praktikabilität, der Qualität, Implementierbarkeit und Akzeptanz (Zweckmäßigkeit). Hinzu kommt häufig die qualitative und quantitative Identifikation von Risiken, die in die Betrachtung aller Dimensionen einbezogen werden müssen. Eine Untersuchung zur Beantwortung einer Frage der Gesundheitsversorgung muss daher in aller Regel gleichzeitig mehrere orthogonale Betrachtungsdimensionen einbeziehen. Die hierfür erforderlichen Daten sind aktuell in keiner Datenquelle vollständig und verfügbar. Bis heute richten sich Forschungsprojekte nicht an prioritären Fragen der Versorgung aus, sondern an den eingeschränkten Möglichkeiten, die durch die jeweils verfügbaren Datenquellen vorgegeben sind. In der Folge betreffen Ergebnisse der Versorgungsforschung in der Regel nur Aspekte eines Versorgungsproblems. In Studien zur Wirksamkeit von neuen Versorgungskonzepten fehlt bspw. häufig eine umfassende gesundheitsökonomische Kosten-Nutzwert-Betrachtung, eine Untersuchung der Auswirkungen einer Maßnahme auf weitere Bestandteile des Versorgungssystems, die Analyse der Praktikabilität der Einführung und Umsetzung und der Akzeptanz durch die Patienten und Leistungserbringer. Die limitierte Datenlage limitiert somit die Möglichkeiten der Versorgungsforschung, und verhindert, behindert oder verzögert die Entstehung von Evidenz im Versorgungssystem. Für die Umsetzung eines evidenzbasierten, "lernenden" und datengestützten Gesundheitssystems müssen bisher unverknüpfte personenbezogene Daten aus verschiedenen Domänen zusammengeführt werden. Nur auf dieser Grundlage kann die Gesundheitsversorgung zielgerichtet verbessert werden.

Daten der GKV spielen hier eine zentrale Rolle. Ihre Vollständigkeit bezogen auf alle GKV-Versicherten bildet das Versorgungsgeschehen sektorübergreifend und bevölkerungsbezogen ab und ermöglicht dadurch eine unselektierte Betrachtung auf Systemebene. Wesentliche Einschränkungen betreffen die Ausrichtung auf die Kostenerstattung. Gesundheitsrelevante patientenindividuelle Faktoren wie Risikofaktoren, Expositionen, aber auch Lebensqualität, Präferenzen, Gesundheitsverhalten und -kompetenz werden nicht erfasst. In Kohortenstudien mit teilweise umfangreicher Primärdatenerhebung von den Studienteilnehmende können diese Faktoren mit adäquater Methodik und entsprechendem Aufwand in hoher Qualität erfasst werden. Da hier der Studienteilnehmer bzw. sein Umfeld die Informationsquellen sind, ist die Erfassung des Versorgungsgeschehens, Inanspruchnahme des Gesundheitssystems, daraus resultierende Kosten in Kohortenstudien in der Regel nicht valide erhebbar. Dies wird deutlich am hier zugrundeliegenden Beispiel der NAPKON Kohorten. Für ein rationales, evidenzgeleitetes Pandemiemanagement sind valide Daten zu Risikokonstellationen, der klinischen Behandlung, dem Monitoring des Verlaufs der Erkrankung und Komplikationen sowie die Patientenerfahrung kombiniert werden mit systembezogenen (Kapazitäten, Organisation, Performanz, Zeitabläufe) und gesundheitsökonomischen Daten (Kosten von

Interventionen und Nicht-Interventionen, Effekte der Leistungserbringung in unterschiedlichen Sektoren, Auswirkungen von diagnostischen Algorithmen und Behandlungsleitlinien auf das Gesamtsystem (z. B. Lockdown, Kapazitätsverschiebungen im stationären Bereich, Schwerpunktbildungen etc.).

Diese Zusammenführung von primär erhobenen Daten (hier: in Kohorten), mit versorgungsnahen Daten aus unterschiedlichen Quellen (hier: zunächst Abrechnungsdaten der GKV) war in der Vergangenheit nur punktuell und mit großem zeitlichen und kommunikativem Aufwand möglich. Durch den Aufbau des FDZ würden wesentliche Barrieren adressiert: eine vollständige, kontinuierliche, zeitnahe Zusammenführung der Daten aus allen gesetzlichen Krankenkassen bilden das Versorgungsgeschehen aller GKV-Versicherten und damit für über 90% der Bevölkerung vollständig ab und schaffen dadurch eine hohe Repräsentativität. Die Dokumentation der Kassenkennzeichen ermöglicht die methodische Abbildung der im System vorhandenen systematischen Heterogenität durch regionale Unterschiede und Selektionseffekte in den Versichertenpopulationen einzelner Krankenkassen (insbesondere durch adäquate Multi-Level-Analysen etwa zur Unterscheidung von versorger- und patientenseitigen Risikofaktoren). Die Vereinheitlichung und Definition der Variablen und ihrer jeweils möglichen Ausprägungen und deren Kategorisierungen ermöglichen valide Aussagen auf der Ebene der Gesamtbevölkerung der GKV-Versicherten. Eine standardisierte Beantragung des Zugangs und ein transparentes Verfahren zur Prüfung der Berechtigung der Antragsteller senken derzeitige bürokratische und organisatorische Hürden, erhöhen die Planbarkeit und beseitigen die aktuell stellenweise vorhandene Interessenleitung einzelner Krankenkassen bei der Gewährung des Datenzugangs. Eine lückenlose Dokumentation der verwendeten Analysepopulation, Variablen und Ausprägungen sowie der angewandten Programme und Algorithmen erhöhen wesentlich die Reproduzierbarkeit und wissenschaftliche Qualität der Forschung und verbessern damit die Interpretier- und Belastbarkeit der Ergebnisse.

Zukünftig sollte die Verknüpfbarkeit mit weiteren wesentlichen Datenquellen im FDZ sichergestellt werden. Hierzu gehören die personenbezogenen Daten aus Registern, insbesondere der Landeskrebsregister nach § 65 c SGB V und weiterer krankheitsbezogener Register (vergl. hierzu [11]). Relevante versorgungsrelevante Daten liegen bspw. beim Institut für Arbeitsmarkt und Berufsforschung (IAB) der Deutschen Rentenversicherung Bund (lebenslange Berufsbiographie der Versicherten), beim Robert-Koch-Institut (RKI) (z. B. Surveys im Rahmen der Gesundheitsberichterstattung, regelmäßige Surveillance-Erhebungen bspw. aus Panels, zentrale Erfassung bspw. von meldepflichtigen Infektionserkrankungen), beim Paul-Ehrlich-Institut (PEI) (z. B. Hämo- und Gewebevigilanz), beim Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) (zentrales Meldesystem für Arzneimittel-Nebenwirkungen) und beim Statistischen Bundesamt. Zukünftig werden die Behandlungsdaten aus den Datenintegrationszentren der Universitätskliniken vorliegen und

durch den erheblichen personellen Aufbau und neue Aufgaben die Gesundheitsämter sowohl wesentliche Datennutzer- als auch Lieferanten von versorgungsnahen Daten werden.

Fragen der Pandemic Preparedness sind zentraler Bestandteil aktueller und zukünftiger Forschungsaktivitäten in der Medizin und im öffentlichen Gesundheitswesen, sowohl in NUM-Projekten wie NAPKON und egePAN Unimed [12] als auch weit darüber hinaus. Die Vorhaltung zentraler Daten und eines zentralen Verfahrens im FDZ zur einfachen und zeitnahen Verknüpfung von GKV-Routinedaten mit Primärdaten, aber auch mit anderen Datensätzen, hätte ein enormes Potenzial, die unverzichtbare Datengrundlage für informierte politische Entscheidungen im Sinne einer evidenzbasierten Pandemiesteuerung schnell verfügbar zu machen. Um die optimale Ausschöpfung der Potenziale der FDZ Infrastruktur, nämlich deren Rasche Aktivierbarkeit (z. B. durch regulatorische fast track Verfahren), der verlässlichen Planbarkeit (z. B. durch standardisierte Datensatzbeschreibungen) und eine hohe Skalierbarkeit (im Sinne der Bereitstellung wachsender Mengen aufbereiteter qualitätsgesicherter verknüpfungsbereiter Daten) sicherzustellen, sollte die Umsetzung von Anfang wissenschaftlich begleitet und zielbezogen evaluiert werden.

Das Ziel muss ein integriertes zentrales Gesundheits-Daten-Ökosystem in Deutschland sein, in der Forschung, Monitoring, Alert-, Informations- und Management-Funktionen zentral unterstützt werden können. In diesem System sollten deskriptive und analytische Versorgungsforschung, Wirksamkeitsstudien (z. B. durch target trial emulation, vergl. [16]), Studien zur Pharmakoepidemiologie, zur Leitlinien-Adhärenz, und zur Qualität einzelner Versorgungsleistungen unterstützt werden. Daneben kann durch vorgegebene spezifische Auswahl von versorgungsnahen Daten aus diesem Ökosystem Kontrollgruppen für auf Primärdaten basierende klinische Studien und Implementierungsstudien gebildet werden. In Verbindung mit geeigneten Primärdaten können versorgungsnahen Daten aus dem Gesundheitsdatenökosystem zur Ursachenforschung und zur Präventionsforschung und Public Health Forschung herangezogen werden. Angesichts der großen Akzeptanz der Gesundheitsforschung auf der Basis von versorgungsnahen Daten in der Bevölkerung sollte es möglich sein, Patientinnen und Patienten gezielt zur Teilnahme an Primärdatenstudien einzuladen. Diese Möglichkeit wurde aktuell in mehreren Landeskrebsregistergesetzen bereits gesetzlich verankert (z. B. Mecklenburg-Vorpommern, Rheinland-Pfalz). Das FDZ könnte auch bei europäischen Initiativen wie dem Aufbau des European Health Data Space eine führende Rolle übernehmen.

2.2 Datenlinkage

Datenlinkage ist von großer Relevanz für die wissenschaftliche Forschung. Ein personenspezifisches Linkage ist Voraussetzung dafür, die Stärken der jeweiligen Einzeldatenquellen miteinander zu kombinieren und dabei zugleich bestimmte Schwächen ebendieser Quellen zu kompensieren. So können etwa in den GKV-Daten langfristige

sektorenübergreifende Erkrankungsverläufe abgebildet werden, welche in den NAPKON-Daten nicht erfasst werden konnten. Umgekehrt werden etwa in der NAPKON-POP-Kohorte der akute Erkrankungsverlauf von Patientinnen und Patienten mit überstandener SARS-CoV-2-Infektion detailliert retrospektiv erfasst und um eine prospektive Biomaterialsammlung sowie die Erfassung der Langzeitfolgen und Komorbiditäten ergänzt, welche in den GKV-Daten nicht oder nicht mit diesem Detailgrad abgebildet sind. Die Verknüpfung von GKV-Routinedaten mit NAPKON-Daten bietet damit die einzigartige Möglichkeit, ein umfassendes Bild der Gesundheitsversorgung der NAPKON-Studienteilnehmende vor und nach der eigenen Sars-Cov2-Infektion sowie vor und nach der Pandemie zu zeichnen. Insbesondere können (Ko-)Morbidity, Medikamente, Verordnungen von Heil- und Hilfsmitteln, stationäre Aufenthalte und ambulante Behandlungen über einen längeren Zeitraum abgebildet werden. Anhand der klinischen Daten wiederum können erweiterte Risikoadjustierungen zu einem vollständigeren Bild führen, als dies allein anhand der GKV-Daten möglich wäre. Durch die Erfassung von Bioproben sind gleichzeitig Untersuchungen zu den pathophysiologischen Grundlagen der Sars-CoV-2-Infektion und der COVID-19 Erkrankung möglich. Wir gehen davon aus, dass FDZ-Daten für Forschungszwecke zukünftig häufig mit einer oder mitunter auch mehreren weiteren Datenquellen verknüpft werden. Vor dem Hintergrund des Broad Consent und dem darin enthaltenen Modul GKV-Daten gehen wir von einer besonders häufigen Indikation für consent-basiertes Linkage zwischen Klinikdaten und GKV-Routinedaten aus.

Allerdings umfasst der Begriff Datenlinkage eine Reihe von Vorgehensweisen, die nachfolgend voneinander abgegrenzt werden. Dabei wird insbesondere das im vorliegenden Gutachten betrachtete Datenlinkage methodisch eingehend vorgestellt. Zunächst ist zwischen dem Linkage auf Individualdaten-Ebene und auf Aggregationsebene zu unterscheiden, wobei hier grundsätzlich nur Daten mit Personenbezug berücksichtigt werden. Bei Individualdaten-Linkage werden Angaben, welche spezifisch für einzelne Individuen sind, über verschiedene Datenquellen (z. B. Diagnose einer Erkrankung für eine Patientin bzw. einen Patienten aus Datenquelle A und Angaben zur Behandlung der Erkrankung bei dieser Patientin bzw. diesem Patienten aus Datenquelle B) so verknüpft, dass diese auch nach Verknüpfung wieder individuell zuzuordnen sind. Diese Art der Verknüpfung ist Gegenstand des vorliegenden Gutachtens und wird nachfolgend weiter betrachtet. Demgegenüber steht die Verknüpfung auf Aggregatebene (ökologische Verknüpfung). Hier werden Angaben verknüpft, welche für eine ganze Gruppe von Individuen gleichermaßen gelten (z. B. Durchschnittseinkommen in einer bestimmten Region mit Eigenschaften dieser Region). Diese Art der Verknüpfung ist nicht Gegenstand des vorliegenden Gutachtens und wird daher nachfolgend nicht weiter betrachtet.

Beim Linkage auf Individualdatenebene werden Angaben zur selben Person aus den unterschiedlichen Datenquellen einander zugeordnet. Diese Zuordnung erfolgt durch den Abgleich von Schlüsselvariablen (Identifikatoren), welche jeweils in beiden zu verknüpfenden Datenquellen vorhanden sein müssen. Werden Daten aus mehreren Datenquellen verlinkt, so

kann das Linkage auch über jeweils unterschiedliche aber paarweise vorhandene Identifikatoren je nach Verfügbarkeit erfolgen. Die Zuordnung im Linkage ist aber keinesfalls gleichzusetzen mit einer Reidentifikation im datenschutzrechtlichen Sinne. Die nachfolgenden methodischen Erläuterungen stellen Inhalte aus [25] komprimiert dar, ohne diese jeweils neu zu referenzieren.

Im Linkage sind je nach Datenverfügbarkeit, Datenqualität und Nutzungszweck Abstufungen hinsichtlich der Genauigkeit bzw. Fehlertoleranz der Zuordnung möglich. Ein exaktes Linkage lässt keinerlei Fehlertoleranz zu, es werden also nur Daten miteinander verknüpft, deren Schlüsselvariablen exakt übereinstimmen. Dieses Verfahren ist besonders geeignet bei Nutzung eines eindeutigen Schlüssels oder Identifikators, welcher eine geringe Fehleranfälligkeit besitzt. Im vorliegenden Kontext bietet sich die Krankenversichertennummer (KVNR) an, da diese mit Hilfe der enthaltenen Prüfziffer auf Gültigkeit geprüft werden kann. Klartextangaben wie Name und Vorname von Personen sind hingegen ein Beispiel für fehleranfällige Schlüsselvariablen, da Namensänderungen z. B. in Folge von Eheschließungen oder leicht unterschiedliche Schreibweisen des Namens derselben Person in den verschiedenen Datenquellen die Verknüpfung im exakten Linkage verhindern würden. Hier sind fehlertolerante Linkageverfahren vorzuziehen. Es gibt zahlreiche Methoden, Datenlinkage mit einer Fehlertoleranz durchzuführen. Hier sind insbesondere das regelbasierte, das distanzbasierte und das probabilistische Linkage zu nennen, welche in [25] detailliert erläutert werden.

Im vorliegenden Szenario des auf einer informierten Einwilligung basierenden (consentbasierten) Linkage von NAPKON- und GKV-Daten ist davon auszugehen, dass für alle Studienteilnehmende mit Einwilligung die KVNR sowie Namen, Geburtsdaten und Adressen im Klartext vorliegen. Mit diesen Angaben ist ein exaktes Linkage möglich, sodass dieses Szenario nachfolgend weiter betrachtet wird. Für die Verknüpfung anderer Datenquellen ist die Verfügbarkeit der KVNR nicht immer sichergestellt, beispielsweise wurde die KVNR in klinischen Krebsregistern in den Jahren vor Umsetzung des Krebsfrüherkennungs- und -registergesetz (KFRG) nicht systematisch erhoben. In diesen Fällen muss ein fehlertolerantes Linkage auf Basis anderer Schlüsselvariablen (z. B. Name, Geschlecht, Geburtsdatum, Wohnadresse) umgesetzt werden, wobei die Fehlertoleranz je nach Anwendungszweck variieren kann. Für deskriptive epidemiologische Untersuchungen auf der Ebene großer Populationen (z. B. Bundesländer) sind anteilig mehr Zuordnungsfehler tolerierbar als für Fragestellungen im Bereich der Qualitätssicherung, der Patientensicherheitsforschung (Pharmakoepidemiologie) oder zur Analyse der Leitliniengerechtigkeit der Versorgung in spezifischen Patientengruppen, die weitgehend oder vollständig zuordnungsfehlerfrei erfolgen müssen. Auch in vielen krankheitsbezogenen Registern ist die KVNR nicht erfasst [11]. Um für die in diesen Registern vorhandenen Patientinnen und Patienten Versorgungsdaten aus den GKVn zu ergänzen, dürfen möglichst keine Fehlzuordnungen vorkommen. Dies gilt in beiden Richtungen- weder dürfen die GKV-Daten mehrerer Versicherter im Zeitverlauf

fälschlich einem Kohortenteilnehmenden zugeordnet werden (Homonymfehler), noch dürfen Daten eines GKV-Versicherten fälschlich mehreren Kohortenteilnehmenden zugeordnet werden (Synonymfehler).

Wenn Fehlzusammenhänge nicht weitestgehend ausgeschlossen werden können, sind analytische Fragestellungen, bei denen es auf Unterschiede zwischen definierten und nicht sehr großen Gruppen ankommt, nur eingeschränkt möglich. Gänzlich kritisch wird es bei Anwendungen zur Identifikation sehr seltener, schwerwiegender Ereignisse (critical incidents), bspw. in der Patientensicherheitsforschung. Hier können fehlerhafte Zuordnungen der GKV-Daten zu Daten aus anderen Quellen sowohl zum Übersehen relevanter Zusammenhänge (z. B. Sinusvenenthrombosen nach Corona-Impfung mit dem Impfstoff von Astra Zeneca, vergl. [12]) oder zu einer fälschlichen Zuordnung seltener, aber erheblicher Nebenwirkungen und Komplikationen zu klinischen Interventionen führen, die tatsächlich ungefährlich oder sogar nützlich sind. Hier ist zu beachten, dass in Deutschland kein anderes System existiert, in dem schwerwiegende Nebenwirkungen und Komplikationen auf der Bevölkerungsebene vollständig erfasst werden. Vor allem aufgrund ihrer begrenzten Teilnehmeranzahl, aber auch in Bezug auf die zeitliche Beschränkung können seltene und/oder spät auftretende Nebenwirkungen und Komplikationen neuer therapeutischer Interventionen nicht im Rahmen der Zulassungsstudien erfasst werden. Die GKV-Daten stellen hier die absehbar einzige Ressource für die Verbesserung der Patientensicherheit dar und deren Nutzung ist deshalb unverzichtbar. Für alle letztgenannten Anwendungen ist eine besonders geringe Fehlerrate notwendige Bedingung zur Erreichung des Forschungszieles.

Für das Linkage können die Schlüsselvariablen im Klartext oder verschlüsselt genutzt werden. Beim Klartextabgleich können auch bei Nichtübereinstimmung hinreichende Ähnlichkeiten der Schlüsselvariablen für eine Verknüpfung genutzt werden, während beim Abgleich verschlüsselter Information nur exakte Übereinstimmungen zu Verknüpfungen führen, da eine Interpretation der verschlüsselten Werte nicht mehr möglich ist. Für verschlüsselte Schlüsselvariablen sind damit Linkageverfahren mit Fehlertoleranz nur eingeschränkt und mit erhöhtem Aufwand umsetzbar. Zudem ist für verschlüsselte identifizierende Angaben eine gezielte Rückfrage durch die umsetzende Stelle im Linkageprozess für Zweifelsfälle ausgeschlossen, wodurch Fehler im Linkageprozess deutlich schwerer oder gar nicht mehr aufgedeckt werden können. Unter Beachtung dieser Limitationen konnten aber dennoch bereits erfolgreich Linkageprozesse inklusive Verschlüsselung realisiert werden [26]. Hierbei ist aber zu beachten, dass für diese verschlüsselungsbasierten Verfahren in der Regel keine besonders geringe Fehlerrate sichergestellt werden kann, weshalb diese nur für Nutzungszwecke mit weniger hohen Anforderungen hinsichtlich der Fehlerrate angewendet werden sollten.

Werden personenindividuelle Daten verschiedener Datenhalter auf Individualebene verknüpft, bestehen besondere Datenschutzerfordernisse. Diese sind insbesondere durch

die Rechtgrundlagen der jeweiligen Datenquellen vorgegeben (gesetzliche Grundlage und/oder informierte Einwilligung). Zudem gelten die Regelungen der DSGVO und entsprechende Bundes- und Landes-Datenschutzgesetze. Um diesen Anforderungen auch im Linkageprozess gerecht zu werden, muss in der Regel für den Linkageprozess ein Trennungsgebot eingehalten werden. Konkret muss sichergestellt sein, dass keine am Linkage-Prozess beteiligte Stelle sowohl die verknüpften Forschungsdaten als auch verknüpfte identifizierende Angaben aus dem Identitätsabgleich erhalten darf (gemäß § 48 Abs. 2 BDSG). Dies erfordert eine Struktur der Datenflüsse mit Einbeziehung einer Treuhandstelle (synonym Vertrauensstelle), welche den Identitätsabgleich vornimmt, aber keine Kenntnis über die Forschungsdaten hat [27-32]. Beispielhaft ist dies in den unter Kapiteln 1.3 und 4.2 beschriebenen Datenflüssen ersichtlich.

Eine Umsetzung von Datenlinkageprozessen ist je nach Komplexität des jeweiligen Verfahrens mit Eigenentwicklungen basierend auf Datenbanksystemen (z. B. SQL-basierte Systeme) bzw. Statistikprogrammen (z. B. R, SAS, Stata) oder mittels spezifischer Software-Tools möglich. Eine Übersicht spezifischer Software-Tools für das Datenlinkage ist beispielsweise in [25] gegeben.

Der Erfolg des Linkage ist aber nicht nur von der technischen (Datenfluss) und methodischen (Linkageverfahren) Ausgestaltung abhängig. Wichtig ist ebenso, dass die zu verknüpfenden Datenquellen eine hinreichende Überlappung hinsichtlich der Schlüsselvariablen zum Identitätsabgleich sowie der zu verknüpfenden Kohorte(n) aufweisen. So kann selbst bei Verknüpfung großer Datenquellen die Schnittmenge klein oder sogar leer sein, wenn die zu verknüpfende(n) Kohorte(n) beispielsweise

- aus nicht oder nicht komplett überlappenden Regionen stammen,
- nicht oder nur teilweise überlappende Einschränkungen hinsichtlich der versichernden Krankenkassen haben,
- aus nicht oder nur teilweise überlappenden Zeiträumen stammen, etwa weil sich die gesetzlichen Datenlöschfristen der Datenquellen unterscheiden (vgl. [26, 33]).

Ebenso kann auch eine mangelnde Vollzähligkeit der jeweiligen Einzeldatenquellen den Linkage-Erfolg einschränken. Sofern aber die sonstigen Charakteristika der zu verknüpfenden Kohorte in beiden Datenquellen identisch eingeschränkt sind, können durch ein gutes Datenlinkage und die Bewertung des Anteils an nicht erfolgreich verknüpften Individuen wertvolle Aussagen zur Vollzähligkeit gewonnen werden. Eine belastbare Quantifizierung der Vollzähligkeit ist allerdings erst dann möglich, wenn eine der Datenquellen entweder nahezu vollzählig ist oder plausibel angenommen werden kann, dass die zur Vollzähligkeit fehlenden Individuen keiner systematischen Selektion unterliegen (Missing Completely at Random, vgl. [34]). Hierbei ist darüber hinaus zu berücksichtigen, dass bei Linkageverfahren mit exakter Übereinstimmung der Schlüsselvariable(n) die Anzahl an erfolgreichen Verknüpfungen sinkt,

wenn entweder die Schlüsselvariablen (teilweise) fehlerbehaftet (z. B. Zahlendreher in den Identifikatoren) sind oder zu viele Schlüsselvariablen genutzt werden (vgl. [26, 33]). Ebenso kann aber auch die Nichtverfügbarkeit relevanter Schlüsselvariablen bzw. mangelnde Vollständigkeit der Angaben in diesen Schlüsselvariablen den Linkageerfolg verringern - etwa die unvollständige Dokumentation der KVNR in historischen KKR-Daten vor Einführung des KFRG, welche im Datenlinkage gegebenenfalls durch andere, weniger genaue Schlüsselvariablen kompensiert werden muss.

Wie bereits betont, ist eine sehr hohe Qualität des Linkage eine zentrale Voraussetzung für die Durchführung der Mehrheit der Projekte der Gesundheitsforschung. Bestehen Zweifel an der korrekten Zuordnung von Daten aus verschiedenen Datenquellen zu einer Person, schränkt dies die Aussagekraft der Ergebnisse erheblich ein. In solchen Zweifelsfällen muss zuweilen in einer Treuhandstelle auf die identifizierenden Klartextangaben zurückgegriffen werden, um die Korrektheit des Linkage zu prüfen und insbesondere falsche Zuordnungen (Homonym- und Synonymfehler) zu korrigieren. Nur so lässt sich ein personenbezogenes Linkage mit Minimierung der fehlerhaften Zuordnungen erreichen.

Insbesondere die bei Kontrollnummern eingesetzten Hashverfahren sind anfällig für unterschiedlichen Schreibweisen der Klartextangaben und führen in der Folge zu Problemen beim Linkage. Unabhängig von dem angewandten Linkageverfahren ist deshalb standardmäßig ein so genanntes Preprocessing durchzuführen, in dem eine Harmonisierung der als Quasi-Identifikatoren bezeichneten Klartextangaben erfolgt.

In Zweifelsfällen ist der Abgleich auf Klartextangaben immer einem Linkageverfahren auf Basis von Kryptoelementen (beispielsweise Kontrollnummern oder Bloomfilter) vorzuziehen. So ist beispielsweise im Landeskrebsregistergesetz im Land Nordrhein Westfalen (LKRK NRW) die Entschlüsselung von Identitäts-Chiffren zulässig, beispielsweise bei so genannten Doppelverdachtsfälle (Synonym- und Homonymfehler) oder Death Certificate Only-Fälle (DCO-Fälle), zu denen allein die Todesfallmeldung vorliegt [35].

Zusammenfassend kann festgehalten werden: Eine One-Size-Fits-All-Lösung für individualdatenbasiertes Datenlinkage gibt es nicht, es muss vielmehr für jede Datensituation individuell entschieden werden, welches Linkageverfahren mit welchen Schlüsselvariablen genutzt wird. Zudem beeinflusst die Qualität der Datenaufbereitung und -harmonisierung deutlich die Qualität des Datenlinkage. Nicht zuletzt spielt bei den sehr großen Datensätzen im FDZ die Performanz der Softwarelösungen eine Rolle. Einige in der Forschung an kleineren Datensätzen verwendete Software-Lösungen sind konzeptbedingt für die Anwendung im FDZ nicht geeignet [36]. Generelle Empfehlungen bezüglich der beim Datenlinkage zu beachtenden Aspekten sind unter [37] zu finden.

Zur Messung des Linkageerfolgs gibt es mehrere Aspekte zu berücksichtigen. Zunächst ist zu bewerten, welchen Anteil der betrachteten (Teil-)Population die zu verknüpfenden Datenquellen jeweils einzeln abbilden und welchen Überdeckungs-/Überlappungsgrad die Schnittmenge dieser Datenquellen aufweist. Hierfür sind oft summarische Vergleiche mit publizierten bevölkerungsbasierten Statistiken sowie konzeptionelle Überlegungen zielführend. Weiterhin ist zu bewerten, wie viele Verknüpfungen innerhalb einer Kohorte mit identischen Einschränkungen bezüglich aller verknüpften Datenquellen erzielt werden und welcher Anteil dabei ein-eindeutig (zu einer Person einer Datenquelle findet sich genau ein Treffer zu einer Person in der anderen Datenquelle und dies gilt auch umgekehrt) oder mehrdeutig (zu einer Person einer Datenquelle finden sich mehrere Treffer zu unterschiedlichen Personen in der anderen Datenquelle) sind. Diese Angaben sind zudem mit der Gesamtzahl an Personen der wie beschrieben eingeschränkten Kohorte in jeder verknüpften Datenquelle zu vergleichen. Für mehrdeutige Verknüpfungen ist klar, dass Fehlzuordnungen (Homonymfehler) vorgenommen wurden. Aber auch ein-eindeutige Verknüpfungen können auf Fehlzuordnungen basieren. Um über die beschriebenen Kriterien hinaus auch noch valide einschätzen zu können, wie viele der Zuordnungen korrekt waren und wie viele der nichtzugeordneten Personen auch tatsächlich nicht in der jeweils anderen Datenquelle verfügbar waren, muss das Linkageergebnis mit einem sogenannten "Goldstandard" verglichen werden, sofern dieser verfügbar ist. Ein Linkage-"Goldstandard" wäre eine fehlerlose (oder zumindest nahezu fehlerlose) Verknüpfung der Datenquellen. Diese kann durch einen Klartextabgleich auf hinreichend detaillierten identifizierenden Angaben mit Einzelfallprüfung aller Zweifelsfälle oder die Verknüpfung basierend auf einem (nahezu) fehlerfreien eindeutigen Identifikator (auch "unique identifier") repräsentiert werden. So können Kennzahlen zur Verknüpfungsgüte basierend auf dem Verhältnis richtigpositiver (Person wurde korrekt zwischen den Datenquellen zugeordnet), falschpositiver (Homonymfehler), richtignegativer (Person wurde korrekterweise nicht zwischen den Datenquellen zugeordnet), falschnegativer (Synonymfehler) Zuordnungen abgeleitet werden. In der realen Situation sind die Anteile an Homonym- und Synonymfehlern tendenziell umgekehrt proportional voneinander abhängig. Somit ist es methodisch bedingt nicht möglich, beide gleichzeitig zu minimieren. Deshalb sollte immer dann, wenn ein Goldstandard in der Routineverknüpfung verfügbar und mit den vorgegebenen Ressourcen realisierbar ist, dieser auch für das Datenlinkage genutzt werden. In diesem Fall muss dann auch keine Bewertung des Anteils an Homonym- und Synonymfehlern mehr erfolgen.

Das Datenlinkage über einen eindeutigen Identifikator als Schlüsselvariable ist mit einer hohen Trefferquote verbunden, wenn der Identifikator in beiden zu verlinkenden Datenquellen entsprechend gut gepflegt ist. Wenn, wie in NAPKON POP am Standort Würzburg die Krankenversichertennummer im Rahmen der Einwilligung qualitätsgesichert erhoben wird (beispielsweise mit Kontrolle der Prüfziffer), ist das Linkage mit den Daten des FDZ Gesundheit oder der Krankenkassen technisch problemlos möglich. Wenn hingegen, wie in NAPKON POP am Standort Kiel, wo zunächst nur Klartextidentifikatoren des Probanden/der Probandin

erhoben werden, muss die KVNR nachgetragen werden. Bei Kenntnis der versichernden Krankenkasse wäre dies als Anfrage bei der jeweiligen Krankenkasse mit überschaubarem Aufwand möglich. Fehlt hingegen die Angabe der versichernden Krankenkasse ist der Aufwand erheblich, da ein zentrales Stammverzeichnis mit Klartextangaben aller GKV-Versicherten nach aktueller Kenntnis nicht existiert. Als Alternative bietet sich hierbei das Stammverzeichnis der Rentenversicherung (§ 150 Absatz 2 SGB VI) die ihrerseits für die Vergabe der Rentenversicherungsnummer (RVNR) zuständig ist. Aus dieser RVNR wird in der Vertrauensstelle Krankenversicherungsnummer (ITSG, Essen) die KVNR mit Hilfe eines abgestimmten Verfahrens erzeugt wird [38]. Das Stammverzeichnis der Rentenversicherung und die Vergabe der RVNR ist damit das Wurzelverzeichnis bzw. der Ausgangspunkt für die Generierung der KVNR. Die KVNR ist somit als sekundärer Identifikator anzusehen.

Für den Fall, dass in einem Forschungsprojekt nur die Klartextangaben eines Versicherten vorliegen, für die Abfrage und Linkage von Daten jedoch die KVNR bzw. das im FDZ Gesundheit vorliegende hieraus generierte Periodenübergreifende Pseudonym (PüP) benötigt wird, wären unter Nutzung aktuell verfügbarer Strukturen folgende Arbeitsschritte notwendig:

- Übermittlung der Klartextangaben an das Stammverzeichnis der Rentenversicherung
- Bei einem Treffer Rückmeldung der RVNR an den Forscher
- Übermittlung der RVNR an die Vertrauensstelle Krankenversicherungsnummer
- Rückmeldung der zugehörigen KVNR an den Forscher
- Übermittlung der KVNR an die Vertrauensstelle des FDZ Gesundheit (RKI) zusammen mit einem Lieferpseudonym
- Generierung des zugehörigen PüP und Weiterleitung zusammen mit dem Lieferpseudonym an das FDZ Gesundheit

Inwieweit die anzufragenden Stellen für Anfrage im Rahmen von Forschungsprojekten genutzt werden dürfen und ob hier standardisierte Verfahren zur Anwendung kommen können, konnte im Rahmen dieses Gutachtens nicht geklärt werden. Es wird vorgeschlagen, dies im Nachgang dieses Gutachtens zu prüfen und ggfs. auch die technischen und rechtlichen Voraussetzungen dafür zu schaffen.

Alternativ kann die Vertrauensstelle des FDZ Gesundheit um ein Stammverzeichnis aller GKV-Versicherten erweitert werden und selbst die Umsetzung von Klartextangaben in die für das Datenlinkage erforderliche KVNR übernehmen. Auch hierfür sind Strukturen und die notwendige gesetzliche Grundlage zu schaffen (siehe hierzu auch Kapitel 4.1.2).

2.3 Rechtliche Grundlage

2.3.1 Generelle Aspekte der Datennutzung im Kontext des FDZ Gesundheit

Die Nutzung großer Datenmengen im FDZ ohne Informierte Einwilligung jedes einzelnen Versicherten erfordert die Abwägung des Grundrechtes auf Informationelle Selbstbestimmung gegen die Notwendigkeit der Forschung zur Klärung zentraler offener Fragen im Gesundheitssystem. In ihrer "Petersberger Erklärung zur datenschutzkonformen Verarbeitung von Gesundheitsdaten in der wissenschaftlichen Forschung" stellt die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder das Grundrecht auf Informationelle Selbstbestimmung (Datenschutz Grundverordnung, § 8 der EU Grundrechtecharta) lediglich der Freiheit der Forschung (Art. 5 Abs. 2 GG) gegenüber: "Eine gesetzliche Grundlage für die Nutzung von Gesundheitsdaten zu wissenschaftlichen Forschungszwecken muss einen Ausgleich insbesondere zwischen den verfassungsrechtlich geschützten Interessen schaffen: dem Recht der betroffenen Personen auf Kontrolle über ihre Daten (sog. „informationelles Selbstbestimmungsrecht“) einerseits und der Forschungsfreiheit der Wissenschaftler und wissenschaftlichen Einrichtungen andererseits." ([39], S. 5). Im vorliegenden Fall greift diese Abwägung jedoch zu kurz, da es im Kontext des FDZ nicht allein um Forschung an sich geht, sondern um die Klärung essentieller offener Fragen im Versorgungssystem, die direkt oder indirekt die Qualität, Effektivität und Effizienz der Patientenversorgung betreffen. Neben der Freiheit der Wissenschaft sind hier Aspekte der Qualitätssicherung von Behandlungen im SGB beachtlich (Art. 74 Abs. 1 Nr. 12 u. 19 GG). Beim Monitoring seltener Komplikationen von Therapien und anderen Interventionen geht es um die Patientensicherheit, und mithin das Recht auf Leben und körperliche Unversehrtheit (Art. 2 Abs. 2 GG). Bei der Parametrisierung und der Umsetzung des Gesundheitssystems in Deutschland, der dazu erforderlichen Institutionen, deren Verteilung und Zugänge bis hin zur Auslegung, regionalen Konzentration, Leistungsumfang, Verfügbarkeit, Arbeitsteilung und Zugang sind ethische Grundsätze bis hin zu Fragen der Würde des Menschen (Art 1, Abs. 1 GG) involviert.

Eine umfassendere Abwägung hat der Gesetzgeber zuletzt im Gesetz zur Errichtung des Implantateregisters Deutschland (EIRD) vorgenommen. Dort wird in § 26 explizit die Informationelle Selbstbestimmung der Betroffenen eingeschränkt, weil andernfalls der Zweck des Gesetzes nicht erreicht werden könnte ([40], § 26 und Begründung des Gesetzes 2019). Nach Auffassung der Gutachter gelten für zentrale Anwendungen der FDZ-Strukturen (Patientensicherheit, post-Marketing Surveillance seltener schwerwiegender Nebenwirkungen oder Komplikationen, Ermittlung von Ausgangs- bzw. Vergleichszahlen bei Verdacht auf Medikamenten oder Impf-bezogenen Erkrankungen etc.) analoge Überlegungen. Mit gutem Grund sind in 303 a-f. SGB V und in der aktuellen Fassung der DaTraV bspw. keine weitergehenden Maßnahmen zum Ausschluss jedes Re-Identifizierungsrisikos (Verrauschung der Daten, differential privacy, Belassung der Daten bei den jeweiligen Krankenkassen und Beschränkung der Analysen auf verteiltes Rechnen etc.) und auch kein individuelles Widerspruchsrecht der Versicherten vorgesehen. Würden diese Maßnahmen nachträglich eingeführt, könnte das FDZ wesentliche Auswertungsoptionen nicht mehr sicherstellen.

Selbstverständlich stellt diese Abwägung besonders hohe Anforderungen an die Datensicherheit und den Datenschutz im FDZ - und während des gesamten Nutzungsprozesses. Die konsequente Umsetzung der informationellen Gewaltenteilung ist durch die Einrichtung einer externen unabhängigen Vertrauensstelle gewährleistet. Die in diesem Gutachten im Detail beschriebenen Datenflüsse und -prozesse entsprechen vollumfänglich den Empfehlungen der TMF, die im Dialog und in kontinuierlicher Abstimmung mit den Landes und Bundesbeauftragten für den Datenschutz und der Informationsfreiheit entwickelt und regelmäßig aktualisiert werden (zuletzt [22]).

Im engeren Kontext dieses Gutachtens ist die individuelle informierte Einwilligung der Kohortenteilnehmenden der NAPKON-POP Grundlage des Linkage mit Daten des FDZ. Hierfür ist keine Einschränkung der Betroffenenrechte erforderlich - deren vollumfängliche Ausübung wird vielmehr durch das Management der NAPKON-Kohorte durch die NUKLEUS-Dateninfrastruktur sichergestellt.

Die Datennutzung im Rahmen der Wissenschaft unterliegt einem strengen Kodex. Die Qualität und konkrete Umsetzung ist deutschlandweit in breit konsentierten Empfehlungen und Leitlinien kodifiziert [1, 41, 42] und werden regelmäßig aktualisiert [43]. In der Scientific Community der Versorgungsforschenden wird besonders verantwortlich mit Sekundärdaten umgegangen. Fälle von Missbrauch oder schwerwiegende Datenschutzverstöße sind bisher nicht bekannt geworden (d. h. eine faktische Grundlage für eine "Kriminalitätsvermutung" gegen Wissenschaftlerinnen und Wissenschaftler fehlt). Die hohe Verantwortlichkeit und der Qualitätsanspruch beim Umgang mit personenbezogenen Daten ist eine relevante Grundlage der Nutzung der Daten des FDZ. Die Position vieler Datenschutzexpertinnen und -experten (zuletzt in der Petersberger Erklärung, [39]) bezieht dies nicht mit ein. Der Verweis auf Risiken im Kontext kommerzieller Datensammlung und -nutzung wie bspw. durch Amazon oder Google gehen hier ins Leere.

2.3.2 Rechtliche Aspekte des Datenlinkage von VeDa, insbesondere GKV-Routinedaten

In dem hier vorliegenden Gutachten geht es in erster Linie um das consentbasierte Linkage von klinischen Primärdaten mit versorgungsnahen Daten, hier GKV-Routinedaten. Der Proband bzw. die Probandin willigen ein, dass die Daten für Forschungszwecke genutzt sowie für die Beantwortung der Fragestellung verlinkt werden dürfen. Die schriftliche Einwilligung ist hierbei im Sinne eines Vertrags zwischen Forscher/Forscherin und Proband/Probandin zu verstehen und bildet damit die rechtliche Grundlage für die Verarbeitung der personenbezogenen Daten. Die informierte Einwilligung in eine Forschungsnutzung ist eine opt-in-Situation. Der erklärte Wille eines Probanden/einer Probandin darf nicht durch gesetzliche Regelungen eingeschränkt werden („Informationelle Entmündigung“). Die

rechtliche Verbindlichkeit der informierten Einwilligung ist seit langem in verwandten Kontexten etabliert, bspw. für die Teilnahme an Medikamentenstudien.

Im Falle der auf einer informierten Einwilligung basierenden Forschung ist in vielen Fällen das Vorliegen eines positiven Ethikvotums der zuständigen Ethikkommission erforderlich. Häufig erfolgt im Rahmen der ethischen Prüfung auch die datenschutzrechtliche Prüfung des vorzulegenden Datenschutzkonzepts, häufig unter Hinzuziehung des/der zuständigen Datenschutzbeauftragten, beispielsweise der jeweiligen Universität. Die Prüfgegenstände und die jeweiligen rechtlichen Grundlagen sind jedoch unterschiedlich.

Wenn das Forschungsprojekt ein Linkage mit Routinedaten *nicht* vorsieht, ist das Vorliegen einer informierten Einwilligung und das positive Ethikvotum als alleinige Grundlage für die rechtmäßige Verarbeitung der probandenbezogenen Daten ausreichend. Wenn jedoch, wie hier im oben beschriebenen Forschungsszenario, auch Sozialdaten (hier GKV–Routinedaten) genutzt und verlinkt werden sollen, wird regelhaft von den Krankenkassen, bei denen die Routinedaten angefragt werden, ein Bescheid nach §75 SGB X gefordert. Dieser Antrag muss von der jeweiligen Krankenkasse bei der jeweils zuständigen Aufsichtsbehörde gestellt werden. Aufsichtsbehörde für bundesweit agierende Krankenkassen ist das Bundesamt für Soziale Sicherung (BAS) in Bonn, für regional agierende Krankenkassen ist in der Regel das Sozialministerium auf Länderebene zuständig. Der Antrag kann im Namen der Krankenkasse auch von der Forschungseinrichtung gestellt werden. Dies ist insbesondere dann sinnvoll, wenn in einem Forschungsprojekt mehrere Krankenkassen kooperieren. Dem Antrag ist dann jeweils ein One-Pager der beteiligten Krankenkassen beizulegen, der als Vordruck beispielsweise auf den Internetseiten des BAS bereitsteht. Im Rahmen dieses Antrags werden unter anderem die im folgenden aufgelisteten Punkte festgelegt:

- Die Daten erhaltende Stelle beim Forscher
- Art und Umfang der zu übermittelnden Daten
- Dauer der Nutzung und Löschfristen

Die Bearbeitung eines Antrags durch die Aufsichtsbehörde nimmt je nach Umfang des Antrags und Auslastung der Behörden in der Regel mehrere Wochen bis Monate in Anspruch. Eine maximale Frist zur Bearbeitung ist dabei nicht zentral vorgegeben. Erst nach positivem Bescheid durch die Aufsichtsbehörde übermittelt die jeweilige Krankenkasse die vereinbarten Routinedaten an den Forscher. In der Regel werden hierbei zusätzlich bilateral Nutzungsvereinbarungen zwischen der Forschungsinstitution und der Krankenkasse schriftlich fixiert.

Dieses Vorgehen, also die Notwendigkeit eines Antrags nach §75 SGB X sollte bei vorliegender Einwilligung aus unserer Sicht zukünftig entfallen. Wir empfehlen, §75 SGB X entsprechend zu modifizieren. Hintergrund ist, dass bei vorliegender informierter Einwilligung zur

Sozialdatennutzung alle anderen Bedingungen zur Datennutzung aus §75 SGB X (zulässiger Zweck, Interessenabwägung, Datenschutzkonzept) bereits durch andere Institutionen und Behörden (Ethikkommission, Datenschutzbeauftragte der Forschenden Institutionen, Datenschutzbeauftragte der Krankenkasse) geprüft werden. (siehe hierzu Kapitel 3.3 und Kapitel 4).

Wenn in diesem Kapitel von "konsentbasiert" gesprochen wurde, so ist hierbei das Vorliegen einer spezifischen Einwilligungserklärung gemeint. Spezifisch ist diese Einwilligung in dem Sinne, dass die Forschungsfrage in der Patienteninformation im Detail beschrieben wurde und die Patienteneinwilligung die Nutzung der Daten insofern erlaubt, als dass es die Beantwortung dieser Forschungsfrage erfordert. Damit unterscheidet sich eine solche spezifische Einwilligung graduell von dem so genannten Broad Consent, wie dieser im Rahmen der Medizininformatikinitiative formuliert wurde. Dieser Broad Consent vereinbart mit dem Probanden/ mit der Probandin einen etwas größeren und breiteren Forschungsrahmen (beispielsweise eine weiter gefasste Forschungsfrage oder medizinische Forschung allgemein). Wie im spezifischen Fall ist auch im Broad Consent der Zeitraum der Datennutzung bzw. der Datenbeobachtungszeiträume genau benannt (beispielsweise 5 Jahre vor und 5 Jahre nach dem Datum der Einwilligung). Der Broad Consent der Medizininformatikinitiative beinhaltet ein „Kassen-Modul“, in dem ein Proband in die Anforderung und Nutzung seiner Krankenkassendaten einwilligt.

Ein Broad Consent ist Ausgangspunkt für die Erhebung der Studiendaten ggfs. bei Anwendung des Kassenmoduls auch der Einbeziehung von GKV-Routinedaten und deren Integration in eine projektübergreifende Studiendatenbank. Hauptunterschied des Broad Consent im Vergleich zu dem studienspezifischen Konsent wie im Beispiel NAPKON-Krankenkassenmodul ist die Personengruppe, von der der Konsent erhoben wird. Beim Broad Consent sind dies alle Patientinnen und Patienten eines Klinikums, bei Kohortenstudien nur Patienten, die die Einschlusskriterien für die jeweilige Kohorte (hier die NAPKON POP) erfüllen. Spezifische Fragestellungen im thematischen Spektrum, das im Broad Consent definiert ist, lassen sich dann ohne Einholung einer projektspezifischen Einwilligung in jedem Einzelfall auf Basis dieser Studiendatenbank beantworten. Durch transparente übergeordnete Governance-Strukturen, beispielsweise durch ein Use & Access Committee, wird der Antragsprozess für spezifische Auswerteprojekte unterstützt und der projektspezifische Zugang zu den Daten geregelt. Abhängig von der jeweiligen Forschungsfrage und den geplanten Auswertungen kann das Use & Access Committee die Vorlage eines positiven Ethikvotums fordern.

2.4 Rolle der Krankenkassen

Bisher ist ein Linkage von Studiendaten mit GKV-Routinedaten für Forschende in erster Linie durch eine projektspezifische Kooperation mit einer oder mehreren gesetzlichen Krankenkassen möglich. Ambulante Versorgungsdaten können auch über die Kassenärztlichen

Vereinigungen verfügbar gemacht werden. Jedoch bieten die Daten der Krankenkassen die einzige Möglichkeit, die Gesundheitsversorgung sektorenübergreifend abzubilden, wie in dem Beispiel dieses Gutachtens vorgesehen (Linkage NAPKON-Daten mit sektorenübergreifenden GKV-Daten). Die Krankenkassen entscheiden nach Anfrage von Forschenden zur GKV-Datennutzung individuell, ob sie das Forschungsvorhaben unterstützen oder nicht. Explizite Kriterien für die Entscheidung einer Krankenkasse für oder gegen die Unterstützung des Forschungsvorhabens existieren nicht. Weder positive noch negative Entscheidungen werden i. d. R. explizit begründet. Nach eigenen Erfahrungen und Ergebnissen der Experteninterviews im Rahmen der Gutachtenerstellung erfolgt die Bewertung der Krankenkassen anhand der Kriterien:

- Relevanz des Projekts aus Sicht der Krankenkasse,
- Vorerfahrung der anfragenden Forschenden im Umgang mit GKV-Routinedaten und
- Aufwand/Kapazitäten der Krankenkasse für das Datenmanagement und die Datenübermittlung
- Rolle der Krankenkasse (Kooperationspartner beim Forschungsprojekt mit inhaltlicher Beteiligung vs. reiner Datenlieferant)

Es gibt derzeit keine gesetzliche Möglichkeit, die Krankenkassen für die Unterstützung eines Forschungsvorhabens durch Übermittlung von GKV-Routinedaten zu verpflichten. Ein Gremium, an das kassenübergreifend Anfragen für Forschungsprojekte zur Datennutzung gerichtet werden kann, ist ebenfalls nicht vorhanden. Forschende müssen daher i. d. R. mit jeder Krankenkasse einzeln verhandeln. Dies ist gerade für Fragestellungen, für die eine hohe Repräsentativität der Studienpopulation und eine hohe Vollständigkeit der Daten erforderlich sind, extrem aufwändig, nicht planbar und typischerweise auch nicht erfolgreich. Bei Zustimmung einer oder mehrerer Krankenkassen zur Projektunterstützung müssen dann unterschiedliche administrative und regulatorische Verfahren durchlaufen werden. Dies beinhaltet die Prüfung durch die Datenschutzbeauftragten der Forschungseinrichtung und der Krankenkasse, die Prüfung durch die zuständige(n) Ethikkommission(en) und die Prüfung durch die Aufsichtsbehörde(n) der beteiligten Krankenkasse(n). Letzteres beinhaltet die Stellung eines Antrags nach §75 SGB X (Übermittlung von Sozialdaten für die Forschung und Planung) an die oberste Bundes- oder Landesbehörde, die für den Bereich, aus dem die Daten herrühren, zuständig ist. Für die Bearbeitungszeit von Anträgen an Landesministerien mit Zuständigkeit für Soziales (Aufsicht für AOKen) und an das Bundesversicherungsamt (Aufsicht bundesweiter Krankenkassen) besteht keine Frist. Erfahrungsgemäß können Rückmeldung bereits nach Wochen erfolgen, oder aber auch fast ein ganzes Jahr brauchen. Auch dies führt für die Forschenden zu einer fehlenden Planbarkeit von Forschungsprojekten mit Linkage von GKV-Routinedaten. Trotz dieser Herausforderungen finden Untersuchungen mit Linkage von GKV-Routinedaten und anderen Daten zunehmend im Rahmen von Projekten des Innovationsfonds statt. Dies betrifft sowohl Projekte der Förderlinie "Versorgungsforschung" als auch die Evaluation im Rahmen der Förderlinie "Neue Versorgungsformen" des

Innovationsfonds. Typischerweise sind die Krankenkassen in diesen Projekten nicht ausschließlich Datenlieferant, sondern aktiv an der Forschung beteiligt (konzeptionell, methodisch, bei der Interpretation der Ergebnisse). Von Krankenkassen unabhängige Forschung mit sektorenübergreifenden GKV-Routinedaten und deren Linkage mit Primärforschungsdaten findet praktisch nicht statt.

2.5 Barrieren: Erfahrungen aus NAPKON

Die Machbarkeitsstudie im NAPKON-Projekt kann zur Veranschaulichung einiger der Schwierigkeiten dienen, die entstehen, wenn Krankenkassen um ihre Beteiligung an Forschungsprojekten angefragt werden. Das Fallbeispiel veranschaulicht den enormen Aufwand und verdeutlicht zudem die Komplexität der Planung, die vor allem auf die große Heterogenität der Anforderungen an die regulatorischen Prozesse und die Kooperationsbereitschaft einer Vielzahl von Krankenversicherern zurückzuführen ist. Ob ein Projekt, in dem die Einwilligung von Probanden und Probandinnen bereits vorliegen, auch realisiert werden kann, lässt sich derzeit zum Zeitpunkt des Projektbeginns nicht mit hinreichender Sicherheit sagen.

Barrieren im NAPKON Use Case

1. Komplexe und langwierige Koordinierung von Datenanfragen bei Krankenversicherungen mit ungewissem oder negativem Ergebnis
2. Ineffizienzen aufgrund von Problemen und Unsicherheiten bei der Erstellung von abgestimmten regulatorischen Dokumenten (z. B. Einverständniserklärungen), die von allen Beteiligten akzeptiert werden.
3. Ineffizienzen durch unklare und/oder (zu) viele behördliche Zuständigkeiten für Entscheidungsprozesse bei Datenschutz und ethischen Belangen
4. Hohe Anforderungen an technisches, methodisches und inhaltliches Know-How im Hinblick auf die Datenflussarchitektur, Datenmanagement und inhaltlichen Aufbau von Kassendaten und deren Eignung für bestimmte Fragestellungen.

Ad 1) Im NAPKON Projekt stand zunächst die Kontaktaufnahme mit den Krankenkassen auf dem Plan. Es erschien sinnvoll, dass vorrangig Krankenkassen angesprochen werden sollten, bei denen besonders viele Patienten in der Population der Machbarkeitsstudie versichert waren. Der Erstkontakt erfolgte dabei über externe Experten (u.a. aus einer eigens dafür zusammengestellten Expertengruppe), die bereits über persönliche Kontakte zu den Krankenkassen verfügten, vor allem auf Grundlage abgeschlossener oder laufender gemeinsamer Forschungsprojekte. Diese Strategie hatte sich in bisherigen Projekten bewährt. Schließlich konnte jedoch trotz mehrerer direkter Anfragen und erläuternder Gesprächsrunden bisher keine einzige Krankenkasse für die Zusammenarbeit gewonnen werden. Gründe für die Ablehnung waren vor allem die vermutet niedrigen Fallzahlen pro

Krankenkasse sowie fehlende Ressourcen, auch wegen anderer laufender Projekte, insbesondere Projekte aus dem Innovationsfonds.

Ad 2) Bei der Ausarbeitung der Vorhabenbeschreibung, der Studienprotokolle und der regulatorischen Unterlagen des NAPKON Projekts wurden bereits Textmodule implementiert die eine informierte Einwilligung zur Verknüpfung der Daten und zur Rekontaktierung der Teilnehmenden ermöglichen sollten. Im Verlauf der Studie traten diesbezüglich dennoch große Schwierigkeiten auf, die darin bestanden, die Dokumente so zu formulieren, dass sie einer späteren Prüfung der regulatorischen Anforderungen durch verschiedene Gremien und Institutionen wie Ethikkommissionen, Datenschutzbehörden, insbesondere aber der einzelnen angefragten Krankenkassen selbst genügten.

Eine Sichtung der Dokumente durch Sekundärdatenexperten führte zu dem Vorschlag, die Textpassagen der Patienteninformation und -einwilligung zunächst Krankenkassenvertretern vorzulegen, um deren Einschätzung einzuholen. Eine Arbeitsgruppe wurde innerhalb von NAPKON gegründet um die vorbereitenden Prozesse zur Datenverknüpfung zu koordinieren und Sekundärdatenexperten (aus dem Gutachterkreis) einzubinden. Aufgrund der Rückmeldungen der Krankenkassenvertreter wurden die Patienteninformationen präzisiert; dies betraf insbesondere die Informationen zu wissenschaftlichen Fragen sowie zu den Passagen zu den Routinedaten.

Ad 3) An der Datensammlung in den 3 NAPKON Kohorten (SÜP, HAP und POP) sind neben den 36 Universitätskliniken bundesweit eine Vielzahl weiterer Studienzentren beteiligt. Dazu zählen auch nicht- universitäre Kliniken und ambulante Arztpraxen. Dementsprechend sind auch viele behördliche Zuständigkeiten in die Durchführung eines solchen Projektes involviert. Die heterogene Beurteilung der Studienunterlagen durch verschiedene Ethikkommissionen wurde bereits oben beschrieben (siehe Kapitel 1.3). Im konkreten Fall führte die ablehnende Haltung einer einzelnen lokalen Ethikkommission dazu, dass das größte Studienzentrum der NAPKON-HAP Studie keine Patienten in ein Daten-Linkage Projekt einbringen kann (dieses Zentrum macht mit Stand 31.12.2022 über 40% aller HAP-Patienten aus). Von der Weiterverfolgung des Kassendatenlinkages wurde daher für die HAP Studie bis auf Weiteres abgesehen.

Ad 4) Schon im aktuellen Stand der Machbarkeitsstudie werden die technischen Herausforderungen deutlich. Für die Entwicklung der Prozesse des Datenflusses, der Datenverknüpfung, der Datenaufbereitung und -plausibilisierung, sowie der Datenanalyse selbst sind spezifische praktische Erfahrungen für den Umgang mit den beteiligten Institutionen, Krankenkassen und darüber hinaus für eine datenschutzkonforme Verarbeitung der gelieferten Daten unumgänglich. Konkrete Herausforderungen sind dabei z.B. das Aufsetzen einer datenschutzkonformen Datenflussarchitektur und einer sicheren Rechnerumgebung, weiterhin die tiefgehende Kenntnis der Datensätze für die Auswahl der

benötigten Variablen, zusätzlich umfangreiche Erfahrung im Umgang mit verschiedenen Datensatz - Formaten von unterschiedlichen Krankenkassen als Voraussetzung für eine erfolgreiche Harmonisierung der Datensätze und schließlich ein inhaltlich-methodisches Verständnis zur Einschätzung der Eignung der Daten für die Beantwortung konkreter Fragestellungen.

Tabelle 3: Bisheriger Zeitlicher Ablauf der Proof of Concept Studie zum Datenlinkage in NUM NAPKON

Meilenstein/Barriere	Zeitschiene	Erläuterung
Rekrutierungsstart NAPKON mit erstem Sozialdatenmodul in Einwilligung und Patienteninformation für alle 3 Kohorten	Nov 2020 first-patient-in	Beginn Sammlung der KVNR in NAPKON-POP
Treffen der NAPKON AG GKV/Sozialdaten	Q1/2022	Abstimmungsgruppe aus NAPKON und Sozialdatenexperten
Abstimmung Wortlaut Einwilligungserklärung und Pat. Info mit ausgewählten Krankenkassen	bis 02/2022	
Start der Implementierung der Unterlagen in SÜP, HAP und POP	Ab 02/2022 Hierzu jeweils Mitteilung/Amendement an die zuständigen Ethikkommissionen	Für die SÜP rekrutieren 28 Universitätskliniken, 17 nicht-universitäre Krankenhäuser und 23 Hausarztpraxen Patienten, für die HAP und das POP 10 bzw. 3 Universitätskliniken
Vorstellung Proof-of-Concept Studie	13.05.2022 (Sitzung des NUKLEUS Lenkungsausschuss)	Beschluss zur Umsetzung der PoC
Vorstellung des Studienkonzepts der PoC für Krankenkassenvertreter (mit hoher Zahl Versicherter in POP)	31.05.2022 1. Treffen	Nachfolgetreffen 06/2022 mit Kassenvertretern
Ablehnung der Einwilligung einer EK für HAP	08/2022 21.09.2022 erneute Ablehnung	HAP wird nicht erneut Einreichen
Schriftliche Anfrage an Krankenkassen zur Kooperation bzgl. Datenlieferung und Datenverknüpfungen	09/ 2022 (n=5 inkl. Betriebs-KK)	10/2022 Antworten der Kassen: Ablehnung (1) bzw. keine Kooperationszusage (2), keine Antwort (2)
Antrag an die EK Würzburg für Durchführung der PoC Studie zum Datenlinkage	Antrag 10.10.2022 Beratung 25.10.2022	Positives Votum 189/22-sc 27.01.2023 (Würzburg)
Einreichung Use and Access Antrag bei NAPKON	22.11.2022	Bewilligungsschreiben 05.01.2023

Abkürzungen: PoC Proof-of-Concept (Machbarkeit); NAPKON Nationales Pandemie Kohortennetzwerk SÜP Sektorenübergreifende Plattform; HAP Hochauflösende Plattform, POP Populationsbasierte Plattform EK Ethikkommission

Zusammenfassend ist die Beteiligung der Krankenkassen an Forschungsprojekten bislang freiwillig und stark von den Forschungsinhalten und strategischen Schwerpunkten der einzelnen Krankenkassen abhängig (Position des Forschers als Bittsteller). Zudem sinkt die Kooperationsbereitschaft tendenziell mit der Anzahl der bei der jeweiligen Krankenkasse versicherten Personen bzw. ist im konsentbasierten Szenario davon abhängig, wie viele Probanden/Probandinnen bei der jeweiligen Krankenkasse versichert sind. Weitere Probleme entstehen durch ineffiziente regulatorische Prozesse, wegen der großen Zahl der an Entscheidungsprozessen beteiligten Akteure und durch die Unsicherheit in Bezug auf erforderliche Entscheidungen. Bezogen auf technisches Know-how muss erwähnt werden, dass nur bei wenigen Krankenkassen ein eigenes Forschungsmanagement, teilweise in Form von eigenständigen Forschungsinstituten vorhanden ist, die Dateninfrastruktur der Krankenkassen selbst für Forschungsanfragen oft nicht vorhanden ist und für die Vorhaltung einer entsprechenden Infrastruktur gegenwärtig auch keine gesetzliche Verpflichtung besteht. Zusammenfassend sind für die Abrechnungsdaten der gesetzlichen Krankenkassen die FAIR-Prinzipien (findable- Auffindbarkeit, accessible- Zugänglichkeit, interoperable- Interoperabilität, reusable- Wiederverwendbarkeit) zum heutigen Stand nicht umgesetzt. Einschränkungen betreffen aktuell alle FAIR-Dimensionen.

2.6 Die Rolle des FDZ Gesundheit

Mit dem Digitale-Versorgung-Gesetz aus dem Jahre 2019 wurde der Grundstein für das Forschungsdatenzentrum Gesundheit im BfArM (Bonn) gelegt. Wie bereits seine Vorgängerstruktur, die so genannte Datenaufbereitungsstelle (angesiedelt im DIMDI, Köln) wird das FDZ Gesundheit die personenbezogenen (pseudonymisierten) GKV-Routinedaten aller Krankenkassen für Forschungszwecke bereitstellen [13]. Der Produktivbetrieb ist nach aktueller Planung für 2023 geplant.

Das FDZ Gesundheit in seiner beschriebenen Struktur kann die Forschung mit GKV-Routinedaten weiter befördern. Als größter Vorteil ist dabei die Zusammenführung der GKV-Routinedaten aus allen Krankenkassen (n=97, Stand: 1. Jan. 2022) [44] zu nennen. Das FDZ Gesundheit hat das Potenzial, entscheidend zu dem Ausbau der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland beizutragen. Dies ist ethisch und politisch geboten, um die Qualität, Sicherheit und Wirtschaftlichkeit der Gesundheitsversorgung evidenzgeleitet zu erhöhen, neue medizinische Anwendungen zu entwickeln und damit die Gesundheit und Lebensqualität der Bevölkerung zu verbessern [45]. So ist die Auffindbarkeit durch fehlende Versichertenverzeichnisse eingeschränkt, dies gilt insbesondere für

Kassenwechsler. Die Zugänglichkeit der Daten von Versicherten von deren Krankenkassen ist für Forschende auch bei bestehender Einwilligung und Vorliegen eines positiven Ethikvotums erst nach Überwindung weiterer formalen Prozesshürden und Einzelverhandlungen mit Krankenkassen möglich. Die Interoperabilität zwischen verschiedenen Krankenkassen ist limitiert und die Wiederverwendbarkeit allenfalls bei Datenkörpern möglich, die von nur einer Kasse bereitgestellt wurden. Von dem eingeschränkten Zeitraum der Datenhaltung bei den Krankenkassen (5 Jahre retrospektiv) sind alle vier Dimensionen betroffen.

Als zentrale Anlaufstelle für Anfragen zur wissenschaftlichen Forschung an und mit GKV-Routinedaten bietet das FDZ eine offensichtliche Erleichterung für Forschende im Vergleich zum aktuellen Procedere, mehrere parallele Kooperationsanfragen an verschiedene Krankenkassen mit möglicherweise konkurrierenden Eigeninteressen zu stellen. Eine Erwartung an das FDZ Gesundheit aus Sicht der Forschenden ist es, einen unbürokratischen zentralen Datenzugang zu den Versicherten aller gesetzlichen Krankenkassen zu erhalten. Mit dem neuen FDZ Gesundheit entfällt auch die oben beschriebene Einzelfallentscheidung der Krankenkassen, denn das FDZ hält die Daten für die gesamte GKV-Population vor. Insbesondere kleinere Krankenkassen mit wenigen Patienten in der Zielpopulation haben nach bisherigen Erfahrungen sehr zurückhaltend und selten in Verbundprojekten kooperiert. In der Studiendatenbank des FDZ sind zukünftig auch die Daten der Versicherten dieser kleinen Krankenkassen enthalten und können technisch problemlos bereitgestellt werden. Zudem sind die Daten über alle Krankenkassen hinweg harmonisiert. Das zumindest legen die Lieferwege und Datensatzbeschreibungen nahe, die im Rahmen der Datenübermittlung von den Krankenkassen an das FDZ standardisiert wurden. Ein Erfahrungsbericht aus dem FDZ Gesundheit zu der Zusammenführung und Bewertung hinsichtlich des Harmonisierungszustandes bzw. des Aufwands, diesen zu erzeugen, liegen zum jetzigen Zeitpunkt allerdings noch nicht vor.

Mit Start des FDZ Gesundheit stehen die Daten der Beobachtungsjahre 2009 bis 2019 sowie für 2021 zur Verfügung. In einer weiteren Datenlieferung werden die Jahre 2020 sowie 2022 gemeinsam ergänzt und anschließend durch jährliche Nachlieferungen jeweils um ein weiteres Beobachtungsjahr aktualisiert. Der im FDZ Gesundheit zukünftig verfügbare Datenkranz, also der Umfang an Tabellen und Variablen sowie der Merkmalsausprägungen, weist ab dem Beobachtungsjahr 2019 im Vergleich zum Vorgängermodell im DIMDI eine deutlich höhere Detailtiefe auf. So liegen unter anderem Leistungsangaben aus dem ambulanten und stationären Bereich vor, sowie ein Unterscheidungsmerkmal der Leistungserbringer (Praxen mit Facharztzugehörigkeit, stationäre Einrichtungen). Der Datenzuschnitt bis zum Beobachtungsjahr 2018 hingegen orientiert sich an der Datensatzbeschreibung der bisherigen so genannten Datenaufbereitungsstelle im DIMDI, in dem lediglich die Verordnungsdaten vollständig und mit tagesgenauem Datum vorlagen. Ambulante Diagnosen hingegen sind ohne Leistungserbringerbezug enthalten, stationäre Diagnosen liegen nur monatsgenau ohne Liegezeiten vor. Zumindest sind ab dem

Beobachtungsjahr 2016 die Daten der Verstorbenen im Sterbejahr enthalten. Bis 2015 fehlen die Daten der Verstorbenen methodisch bedingt, da sich der frühere Datenpool aus den Morbi-RSA-Daten speiste und die Daten von Verstorbenen im Sterbejahr für das prospektive Morbi-RSA Verfahren nicht notwendig waren. Die beschriebenen Datenbeschränkungen werden durch das neue Lieferverfahren (direkt von den Krankenkassen) und den neuen Datenkranz weitgehend aufgehoben. Allerdings muss darauf hingewiesen werden, dass auch im neuen Datenkranz des FDZ im Hinblick auf die für (Versorgungs-)Forschung relevante Inhalte fehlen. Die nachfolgende Liste im FDZ Gesundheit nicht abgebildeter Angaben zeigt diese Datenlücken an zwei Beispielen auf:

- vollständiges Geburtsdatum
- genaue Wohnortangaben einschließlich Straßen/Hausnummer (wenn auswertungstechnisch begründet)
- Einzelne Kapitel "Spezialversorgung" des Vierten Abschnitts SGB V (§ 115 bis § 122)

Eine abschließende Bewertung kann zum jetzigen Zeitpunkt nicht erfolgen, da nach Auskunft des FDZ Gesundheit das neue Datenmodell bislang noch nicht veröffentlicht wurde und zudem in Abhängigkeit des Zeitpunkts der Datenbereitstellung durch die Krankenkassen aktuell drei Datenmodelle unterschieden werden.

Das neue FDZ Gesundheit will berechtigten Datennutzern (§ 303e SGB V) einen forschungsfreundlichen Datenzugang ermöglichen, der mit den Schlagworten "Virtuelle Analyseräume, Skalierbarkeit, schnellere Berechnung von Ergebnissen, Beantwortung spezifischer Forschungsfragen, zügigere Antragsbearbeitung" umrissen wird [46]. Erfahrungen von Datennutzenden liegen zum jetzigen Zeitpunkt noch nicht vor, da der Produktivbetrieb erst für 2023 in Aussicht gestellt ist.

Das FDZ Gesundheit in seiner jetzigen Konzeption ist als reine Datenplattform geplant. Das FDZ baut hierfür eine Dateninfrastruktur auf, so dass der Forscher/die Forscherin die Daten per personalisiertem und projektspezifischem Fernzugang analysieren kann. Ein Datenexport ist nicht vorgesehen. Ausnahme sind so genannte pseudonymisierte Einzeldatensätze, die dem Forscher auf Antrag bereitgestellt werden und insbesondere der Skripterstellung bzw. Datenplausibilisierung dienen. Exportiert und bereitgestellt werden letztendlich ausschließlich aggregierte Ergebnismengen.

Angesichts der hohen technischen Anforderungen beim Umgang mit den komplexen Datensätzen wie den Routinedaten der Gesetzlichen Krankenversicherungen, wird in vielen Fällen eine unterstützende Beratung durch das FDZ und andere Experten notwendig werden. Diese sollte durch umfangreiche Schulungsmöglichkeiten unterstützt werden. Um das volle Potenzial der FDZ-Daten und eine erfolgreiche wissenschaftliche Nutzung zu gewährleisten,

sollte die Teilnahme der Antragstellenden an einer Basisschulungen für die Datennutzung verpflichtend sein.

2.7 Zukünftige Rollen von Krankenkassen und FDZ Gesundheit

Auch nach Start des Produktivbetriebs des FDZ Gesundheit wird es in Zukunft Forschungsprojekte geben, die sowohl in Kooperation mit Krankenkassen und deren GKV-Routinedaten als auch mit den im FDZ Gesundheit bereitgestellten Datenkörpern durchgeführt werden. Nachfolgend werden Eckpunkte (Rahmenbedingungen) formuliert, mit Hilfe derer zukünftige Forschungsprojekte den für die Fragestellung jeweils geeigneten Zugangsweg festlegen können.

Eckpunkte FDZ Gesundheit

- Bereitstellung von Daten der GKV-Versichertenschaft (ca. 71,5 Mio. Versicherte) über alle Krankenkassen
- Harmonisierte Bereitstellung der Daten im Hinblick auf Datenformate, Benennung der Variablen und Ausprägungen der Feldinhalte. Hierdurch keine aufwändige Aufbereitung notwendig.
- Datenzugang für den Kreis der nutzungsberechtigten Institutionen geregelt
- Antragsverfahren und Datennutzung transparent umgesetzt
- Längere Beobachtungszeiträume möglich, da das FDZ Gesundheit keinen Löschpflichten unterliegt. Je nach benötigtem Datenspektrum greift dieser Punkt erst zu einem späteren Zeitpunkt, da einige Tabellen mit den aktuell erfolgten Lieferungen erstmals in größerer Detailtiefe vorliegen.

Eckpunkte Krankenkasse/Pflegekasse

- Datenlinkage mit externen Daten über Klartextangaben des Versicherten (Nachname, Vorname, Geburtsdatum Wohnort) möglich
- Versicherte können angeschrieben werden, z. B. zur Erhebung weiterer Primärdaten oder Rekrutierung in Studien
- Bereitstellung von Daten der Gesetzlichen Krankenversicherung, die bislang nicht im FDZ Gesundheit enthalten sind (siehe Kapitel 2.5)
- Bereitstellung von Daten der Pflegeversicherung (bislang nicht im FDZ Gesundheit enthalten)
- Bereitstellung zeitlich aktueller Daten möglich: z. B. stationäre Versorgung wenige Tage nach Fallabschluss, Verordnungsdaten wenige Wochen nach Rezepteinreichung
- Verhandlungen mit Krankenkassen über Teilnahme notwendig

- Harmonisierung der von den einzelnen Krankenkassen bereitgestellten Daten teilweise sehr aufwändig
- Löschfristen in den Krankenkassen führen dazu, dass für Studien benötigte längere Beobachtungszeiträume nicht mehr vorliegen

Zu den ersten beiden Punkten muss gesagt werden, dass diese Vorteile zumindest solange gelten, bis eine mit dem FDZ Gesundheit kooperierende Vertrauensstelle existiert, die Klartextangaben der Versicherten vorhält(siehe hierzu auch Kapitel 4.1.2).

Welche Tabellen und Spalten und in welchem zeitlichen Umfang diese Daten zur Beantwortung der Forschungsfrage benötigt werden, lässt sich in der Regel projektspezifisch aus dem Studienplan ableiten. Die hierfür benötigte Datensatzbeschreibung sowie die Begründung für die Spalten sowie die Beobachtungszeiträume sind im Rahmen des positiven Ethikvotums abzustimmen und vom Forscher dem FDZ Gesundheit vorzulegen. Die inhaltlich-methodische Prüfung sollte weiterhin durch die zuständige Ethikkommission übernommen werden, wobei für multizentrische Studien das Prinzip der federführenden Ethikkommission eingeführt werden sollte. Das FDZ Gesundheit prüft lediglich, ob die gesetzlichen Voraussetzungen für die Datennutzung im Einzelfall erfüllt sind und v.a. ob die beantragten Daten für die Forschungsfrage benötigt werden. Eine separate inhaltliche und methodische Prüfung jedes einzelnen Forschungsprojektes durch das FDZ ist nicht erforderlich.

Die Zulässigkeit der Übermittlung von Sozialdaten zu Zwecken der wissenschaftlichen Forschung im Sozialleistungsbereich kann sich aus einer Einwilligung der betroffenen Person ergeben. Art. 9 Abs. 2 lit. a) DSGVO, der die Verarbeitung besonderer Kategorien personenbezogener Daten regelt - worunter auch Sozialdaten fallen -, findet dabei unmittelbar Anwendung. Im nationalen Recht werden in § 67b Abs. 1 S. 2, Abs. 3 SGB X zudem hinsichtlich der Verarbeitung von Sozialdaten bestimmte Anforderungen an die Einwilligung normiert. § 67b Abs. 1 S. 3 SGB X verweist für die Übermittlung von Gesundheitsdaten u.a. auf § 75 SGB X, der eine Übermittlung zu Forschungszwecken auch ohne Einwilligung der betroffenen Person zulässt, wenn die Einholung der Einwilligung unzumutbar ist. Ist eine Einwilligung wirksam erteilt worden, sind jedoch keine weiteren gesetzlichen Tatbestandsvoraussetzungen zu prüfen [47].

Bislang wird von kooperierenden Krankenkassen jedoch die Durchführung eines Verfahrens nach § 75 SGB X für die Bereitstellung der Daten vorausgesetzt und zwar unabhängig davon, ob eine Einwilligungserklärung der Probandinnen und Probanden vorliegt. Insbesondere wird die Genehmigung der obersten Aufsichtsbehörde gefordert. Diese Vorgehensweise entspricht der Vorgabe des § 75 Satz 4: „Die Übermittlung nach Absatz 1 und die weitere Verarbeitung sowie die Übermittlung nach Absatz 2 bedürfen der vorherigen Genehmigung durch die oberste Bundes- oder Landesbehörde“. Nach aktuellen Erfahrungen der Gutachter kann

dieses Verfahren mehrere Monate dauern und erfordert oftmals aufwändige iterative Abstimmungsschritte.

Die Durchführung des Verfahrens nach § 75 SGB X ist jedoch bei Vorliegen einer Einwilligung der betroffenen Person gerade nicht notwendig. Die von den kooperierenden Krankenkassen vorausgesetzte Durchführung des § 75 SGB X-Verfahrens, geht insoweit über das Gesetz hinaus und erschwert die Forschungsdatennutzung erheblich.

Die Gutachter empfehlen daher in Übereinstimmung mit den Empfehlungen der interviewten Ethikerinnen und Ethiker, die Krankenkassen dazu anzuhalten, bei Projekten auf der Basis wirksamer Einwilligungen und Vorlage eines positiven Votums der zuständigen Ethikkommission auf die Durchführung des Verfahrens nach § 75 SGB X zu verzichten oder im Sinne eines „fast lane“-Verfahrens zu verkürzen, da die in dem Paragraphen zu prüfenden Voraussetzungen für die Zulässigkeit der Übermittlung von Sozialdaten wie Nutzungszweck, Interessenabwägung und Datenschutzkonzept bereits durch die zuständige Ethikkommission bewertet wurden.

3 Perspektive und Einordnung der Beteiligten

3.1 Sicht der Krankenkassenvertretenden

Im Rahmen der Gutachtenerstellung fanden zwei jeweils 2-stündige moderierte Expertinnen/Expertendiskussionsrunden statt, an denen jeweils 4 bis 6 Vertretende gesetzlicher Krankenkassen (KK) und Forschende mit Erfahrung in GKV-Datennutzung und Linkage teilnahmen. Wie danken den Expertinnen und Experten (Namen siehe Danksagung) sehr für den offenen Austausch und die vielen wichtigen Empfehlungen, die wir in der Folge nach Themen geordnet zusammenfassen.

Allgemeine Aspekte der GKV-Datenanalyse

Alle KK-Vertretende führten aus, dass die Nachfrage nach GKV-Daten von Forschenden sehr hoch sei und seitens der angefragten KK eine Auswahl getroffen werden müsse, welchen Anfragen nachgekommen werde. Bei allen vertretenen Krankenkassen hänge die Entscheidung der Projektunterstützung durch Datenlieferung davon ab, (1) wie die Relevanz des Projekts (auch und vor allem durch die Krankenkasse selbst) durch die KK bewertet werde, ob (2) die anfragenden Forschende erfahren im Umgang und mit der Auswertung von GKV-Daten seien und (3) ob die KK nur als Datenlieferant oder als Forschungspartner gesehen werde. Projekte, die aus Sicht der KK nicht relevant seien, würden nicht unterstützt. Projekte, bei denen KK nur die Rolle der Datenlieferanten hätten, würden häufig auch nicht unterstützt und besonders in diesen Fällen würden die KK perspektivisch gerne an das FDZ Gesundheit verweisen.

Einheitlich wird von den KK-Vertretende reflektiert, dass sie der hohen Zahl an Anfragen zur Bereitstellung von GKV-Routinedaten nicht nachkommen könnten, weil sie dafür gar nicht die Ressourcen und den Auftrag hätten. Dabei erschwere der Datenschutz die Forschungsunterstützung erheblich. Der Nutzungszweck der GKV-Routinedaten sollte nach einhelliger Auffassung der KK-Vertretende durch den Gesetzgeber auf den Bereich „Forschung“ ausgeweitet werden. Das würde viele der momentan bestehenden Probleme lösen. *(Hinweis: In SGB V §287 findet sich hierzu lediglich eine leistungs- und fallbeziehbare Möglichkeit der Datenanalyse durch die Kassen selbst. Eine Nutzung durch Dritte (Forschende) ist per §75 SGB X auf Antrag möglich.)*

Für "kleine" KK mit Versichertenzahlen im niedrigen 5-stelligen Bereich (z. B. viele BKKen) sei der Aufwand für die Datenaufbereitung und Datenbereitstellung für das FDZ Gesundheit verhältnismäßig groß. Auch gebe es hier anders als bei den großen KK weniger Erfahrungen und Kapazität für die Unterstützung von GKV-Datenanalysen. Aus diesem Grund arbeiteten kleinere KK häufig mit wissenschaftlichen Instituten zusammen, wie bspw. mit dem InGef-Institut.

Die Hauptarbeit für die KK stelle bei Forschungsprojekten nicht die eigentliche Datenbereitstellung dar, sondern es sei sehr viel Beratung durch die KK notwendig, da der Kreis der Forschende mit ausgewiesener Expertise im Umgang mit GKV-Daten begrenzt ist. Entsprechend wurde die Sorge geäußert, dass möglicherweise am FDZ Gesundheit aktuell auch noch zu wenig Erfahrung bei der wissenschaftlichen Nutzung der GKV-Daten vorliege. Die KK-Vertretende betonten einmütig, dass ausreichende Ressourcen in der fachlichen Begleitung von GKV-Datenanalysen dringend notwendig seien. Für das Gelingen eines Projekts sei es unbedingt notwendig, dass methodische Expertise im Umgang mit und bei der Interpretation von GKV-Routinedaten integriert werde, z. B. im Sinne einer beratenden (wissenschaftlichen) Begleitung. Es sei bislang noch unklar, wie das FDZ Gesundheit hier zukünftig aufgestellt ist. Aus KK-Sicht könne eine Verschiebung von Projekten zum FDZ einerseits kasseninterne Ressourcen freischaufeln, andererseits werde befürchtet, dass der Kontakt der KK zu den Wissenschaftlern, der für die KK sehr wichtig sei, gänzlich verloren geht.

Es gebe eine Vielzahl „neuer wissenschaftlicher Institute“, die bislang wenig Erfahrung im Umgang mit GKV-Daten für wissenschaftlich Zwecke besäßen. Dies sei ein Risiko für die Qualität der Projekte. Die Sorge um die Qualität der Evidenz aus GKV-Analysen ohne fachliche Begleitung wurde von den Teilnehmende uneingeschränkt geteilt.

Erwartungen an das FDZ Gesundheit

Insgesamt schätzen die KK-Vertretende das Potenzial des FDZ Gesundheit als sehr hoch ein. Das Glas sei mehr voll als leer. Der Datenkranz sei schon in Stufe 1 sehr detailtief. Stufe 2 werde noch tiefer gehen und es werde sehr herausfordernd für die Kassen, einen

harmonisierten Datensatz zu liefern. Wichtig sei es aus KK-Sicht, praktikable Instrumente und Verfahren für die Forschenden zur Nutzung der FDZ-Daten aufzubauen, sowie begleitende Gremien und Strukturen. Aus Sicht eines KK-Vertreters könne das FDZ Gesundheit, wenn es erfolgreich ist, einen „Big Bang“ in der Versorgungsforschungslandschaft erzeugen. Dazu seien jetzt aber die richtigen Weichen zu stellen. Wichtig sei es nach Übereinstimmung aller Expertinnen und Experten, dass das FDZ Gesundheit von Anfang an hinsichtlich der Prozesse und Nutzbarkeit begleitend evaluiert werde, um möglichst frühzeitig problematische Prozesse zu erkennen und auf Ebene der Verfahren und Prozesse, aber durchaus auch seitens der gesetzlichen Grundlagen entsprechend entgegensteuern zu können. Das FDZ Gesundheit könne nur ein Erfolg werden, wenn klar erkennbar ist, wo die Vorteile und der Mehrwert sowohl für die für Forscher als auch die Kassen liegen.

Spannungsfeld Datenschutz und notwendiger wissenschaftlicher Arbeitsweise mit GKV-Routinedaten

Mehrere KK-Vertretende betonten, dass die Versicherten ihre Daten in der Regel bereitwillig für die Forschung teilen, aber der Datenschutz dem oft entgegenstehe. Es sei entscheidend, dass die Datenschutzregelungen für die wissenschaftliche Nutzung der Daten des FDZ Gesundheit wie auch der Daten einzelner Kassen allgemein angemessen seien. Hier sei es wesentlich, zwischen theoretischer Separierbarkeit und praktischer Identifizierbarkeit zu unterscheiden. Nicht jede Variablenkombination, die selten vorkommt, verursache automatisch eine Identifizierbarkeit. Dieser Sachverhalt müsse unbedingt bei der Datenschutzbewertung berücksichtigt werden.

Projektbeteiligung der KK als Dateneigner auch nach Inbetriebnahme des FDZ Gesundheit

Die KK sind sich einig, dass es auch nach Inbetriebnahme des FDZ Gesundheit Indikationen für die direkte Nutzung der KK-Daten über einzelne KK geben werde. Als Beispiele wurden genannt:

- Projekte, die größeren Datenumfang benötigen als am FDZ vorliegend (zeitlich, Variablen, z. B. Selektivverträge, Pflegedaten)
- Dringende, zeitkritische Analysen (z. B. im Rahmen Evidenzgenerierung bei Pandemie oder Krise)
- Projekte, an denen die KK besonderes Interesse haben und selbst eine aktive Rolle im Projekt einnehmen
- Interventionsstudien
- Grundsätzlich alle Neuen Versorgungsformen des Innovationsfonds (mehr Flexibilität im Projektverlauf, Generierung von Daten aus den Neuen Versorgungsformen, Implementierung der Projektergebnisse durch die KK)
- Versorgungsmanagement (Fragestellungen, die über klassische Versorgungsforschung hinausgehen)

- Je nach Weiterentwicklung des FDZ Gesundheit Studien mit Record-Linkage (*siehe unten ausführlich*)

Aufgrund dieses Bedarfs sollten im Rahmen des Gesundheitsdatennutzungsgesetzes auch Regelungen zur forschungsfreundlichen Umsetzung von GKV-Datenanalysen durch Forschende in direkter Kooperation mit KK getroffen werden. Es gebe eine Vielzahl von Rechtsgrundlagen aus dem SGB, die nicht mehr zeitgemäß sind oder sich gegenseitig widersprechen. Weiteres gemeinsames Fazit war, dass es nicht passieren dürfe, dass durch die Einrichtung des FDZ die Datennutzungsmöglichkeit für die Forschenden bei den Kassen eingeschränkt werden. Die Möglichkeit der direkten Kooperation von Forschenden mit den Kassen müsse erhalten bleiben.

Nach Einschätzung der KK-Vertretenden gebe es auch klare Fragestellungen, die zukünftig primär über ein forschungsfreundliches FDZ Gesundheit umgesetzt werden sollten. Hier sehen sich die Krankenkassen auch klar selbst als Nutzer des FDZ Gesundheit:

- Untersuchungen zu Seltenen Erkrankungen
- Fragestellungen, bei denen die KK-übergreifende Datenvollständigkeit zentral ist, z. B. Versorgungsmonitoring, Fragestellungen der QS
- Kleinräumige Analysen
- Untersuchungen der Patientensicherheit. Hier ist ein vollständiger unselektierter Datensatz entscheidend für die Validität des Ergebnisses. Gerade für solche Untersuchungen wird eine Widerspruchsmöglichkeit als sehr kritisch angesehen. Eine selektionsfreie vollständige Datenbasis ist hierfür vielmehr notwendige Voraussetzung.

Relevanz von Studien mit Record-Linkage

Ein besonderes Interesse liegt sowohl aus Forschungs- als auch aus KK-Perspektive im Linkage von GKV-Routinedaten mit Daten der Klinischen Krebsregister (KKR). Dies müsse auch ohne Einwilligung rechtssicher ermöglicht werden, um die Evidenzlage u. a. für Leitlinienempfehlungen und damit die medizinische Versorgung von krebserkrankten Menschen in Deutschland entscheidend zu verbessern. Problematisch sei hier jedoch die Uneinheitlichkeit der Regelungen zwischen einigen Landeskrebsregistern. Hier sei dringend eine Harmonisierung notwendig, so dass der große Datenschatz, der durch die Verbindung von GKV-Routinedaten und KKR-Daten entstehe, für den medizinischen Fortschritt genutzt werden könne.

Ein weiteres wesentliches Beispiel ist das Linkage von GKV-Routinedaten mit den Daten der MII auf Basis des Broad Consent. Hier willigen rund 70-80% der angefragten Patientinnen und Patienten der Universitätsmedizin in das Linkage der GKV-Routinedaten mit ihren klinischen

Behandlungsdaten ein. Schon nach kurzer Zeit liegen aktuell bereits rund 70.000 Einwilligungen vor, davon gut 50.000 mit Einwilligung für die Verlinkung mit den Krankenkassendaten. Die Ermöglichung des Record-Linkage unabhängig von der Kassenzugehörigkeit ist für die Forscher:innen von größter Bedeutung und verspricht die Generierung von entscheidender neuer Evidenz für die Verbesserung der medizinischen Versorgung in Deutschland. Die Krankenkassenvertretende sehen jedoch auch hier ihre Unterstützung als optional an und möchten sich eine eigene Entscheidung vorbehalten, da ihr gesetzlicher Auftrag derzeit die Bereitstellung von Daten für die Forschung nicht umfasst.

Die Erstellung von Blaupausen für Linkagestudien wird allgemein als sehr sinnvoll erachtet, etwa für eine höhere Modularisierung und Standardisierung Dokumente zur Aufklärung, Einwilligung und zur Datenanforderung.

3.2 Sicht der Forschung

Erwartungen an das FDZ Gesundheit

Auf die Frage „*Was erwarten Sie von FDZ in dieser Ausbaustufe? Wie wird sich die die Forschungslandschaft bei Produktivbetrieb des FDZ verändern?*“ wurde von KK-Vertretenden und Forschenden gleichsam betont, dass die ersten Projekte und Erfahrungen mit dem FDZ Gesundheit entscheidend seien für die spätere Nutzung der Infrastruktur durch die Forschungscommunity. Man dürfe die Fehler aus dem „Datenpool des DIMDI“, insbesondere die langen Bearbeitungszeiten und Prüfprozesse der Ergebnismengen, auf keinen Fall wiederholen. Essenziell sei eine unbürokratische, flexible, forschungsfreundliche Nutzungsmöglichkeit durch die Forschenden (einfache Antragstellung, schnelle Datenbereitstellung, ein iteratives Verfahren aus Analyseschritten und Beurteilung von Zwischenergebnissen zur kontinuierlichen Plausibilisierung) Ein direkter Datenzugriff sei dringend erforderlich mit kurzen Antwortzeiten des Systems. Essenziell sei auch, ein möglichst geringer Zeitverzug zwischen Datenentstehung und Datenbereitstellung. Hier komme beim FDZ Gesundheit zusätzlich zur verkürzten Zeitdauer bis zur Datenlieferung an die KK (ca. 9 Monate bei ambulanten Daten, deutlich weniger bei stationären Daten) die Dauer für die Datenaufbereitung (Änderungen bei den Kassen müssen auch im FDZ berücksichtigt werden) sowie die Dauer des Harmonisierungs- und Plausibilisierungsprozesses der Daten bei der KK-übergreifenden Zusammenstellung hinzu. Beide Schritte seien potenzielle Barrieren und außerdem auch Fehlerquellen. Hinzu kämen administrative Bearbeitungszeiten am FDZ Gesundheit, die hinsichtlich ihres Umfangs bisher nicht eingeschätzt werden können.

Als besonders wichtig wurde einheitlich von den befragten Forschenden herausgehoben, dass für die Forschenden beim FDZ eine Möglichkeit der breiten Nutzung der Daten des FDZ Gesundheit gegeben sein müsse. Es sei bei GKV-Routinedatenanalysen nicht möglich, alles 100%ig vorab festzulegen, sondern die vorab festgelegten Regeln (z. B. im Sinne eines

Studienprotokolls/-plans) sind üblicherweise während der Projektbearbeitung datenbasiert anzupassen und es sind sehr häufig weitere Plausibilisierungen, Sensitivitätsanalysen oder neue Operationalisierungen von Variablen vorzunehmen. Daher sei ein breiter und flexibler Zugriff auf die Datenbasis am FDZ entscheidend, der - beispielsweise zur Plausibilitätsprüfung, Parametrisierung von Modellen und für Sensitivitätsanalysen - auch Queranalysen an Daten/Populationen zulasse, die nicht direkt mit dem beantragten Forschungsthema in Beziehung stehen. Diese flexible, den Sekundärdaten adäquate Vorgehensweise, könne nur per Fernzugriff für die Forschenden praktikabel umgesetzt werden. Alle teilnehmenden Forscherinnen und Forscher bestätigen diesen Aspekt aus langjähriger Erfahrung mit vielen Projekten und sind sich einig, dass zwar 70 bis 90% der Analyseschritte vorgedacht werden könnten, aber dass es praktisch kein Projekt gebe, in dem nicht noch im Verlauf weitere Klärungen und Analysen notwendig würden, um ein qualitativ angemessenes Ergebnis der Forschung zu erzielen. Jedes Sekundärdatenprojekt habe auch explorative Elemente. Analoge Erfahrungen wurden auch im früheren FDZ beim DIMDI gemacht.

Spannungsfeld zwischen Anforderungen durch den Datenschutz und notwendiger wissenschaftlicher Arbeitsweise mit GKV-Routinedaten

Es bestehe hier also offensichtlich ein Spannungsfeld zwischen wissenschaftlich notwendiger Arbeitsweise auf dem Gebiet der Versorgungsforschung mit GKV-Routinedaten einerseits und dem Datenschutz andererseits. Versorgungsforschung unterscheide sich von der gängigen Arbeitsweise in der Epidemiologie und Evidenzbasierter Medizin. Ziel müsse es sein, den Anteil der im Vorhinein nicht exakt planbaren Analyseschritte (siehe oben, je nach Erfahrung der Forschenden und Projekthalte von geschätzt 10% bis 30%) zu verkleinern, bevor der Volldatensatz frei gegeben wird. Lösung könnte auch zunächst ein Teildatensatz sein, der zur Verfügung gestellt wird. Ungeplante Änderungen ergeben sich beispielsweise, wenn Einträge in beantragten Variablen nicht valide oder plausibel sind. Wenn in einer beantragten Variable zu viele fehlende Werte enthalten sind, kann häufig eine neue Hilfsvariable aus anderen (nicht beantragten) Variablen definiert werden, die dann in einem neuen Arbeitsschritt validiert werden muss. Derartige Szenarien kommen erfahrungsgemäß häufig vor und sind nicht vorab planbar. Die Bereitstellung synthetischer Daten wird von den Forschenden hier nicht als zielführend angesehen, auch wenn eine abschließende Bewertung hier noch nicht möglich ist. Es könnte durchaus passieren bzw. muss befürchtet werden, dass in synthetischen Datensätzen ein in den Realdaten vorhandener Zusammenhang durch die angewandten Syntheseverfahren möglicherweise nicht mehr in gleicher Ausprägung enthalten oder durch das Verfahren sogar völlig eliminiert worden ist. Synthetische Datensätze können aus diesem Grund allenfalls Testläufe ersetzen, die eigentlichen Analysen müssen auch auf der vorbereitenden Stufe notwendig auf den Roh-/Realdaten erfolgen.

Belastend und unangemessen sei eine „permanente Kriminalitätsvermutung“ gegenüber den Forschenden. Keinem der Teilnehmer waren Fälle von Missbrauch der Daten bekannt. Die

Orientierung am Ausschluss möglicher Missbräuche ist dennoch derzeitiges Leitbild des Datenschutzes und die aus dieser Orientierung folgenden Maßnahmen machen viele wichtige Projekte schwierig, wenn nicht gar unmöglich. Wie auch die Krankenkassenvertretenden betonen die Forschenden die notwendige Differenzierung zwischen theoretischer Separierbarkeit und praktischer Identifizierbarkeit. Nicht jede Variablenkombination, die selten vorkommt, verursache automatisch eine Identifizierbarkeit. Zur Identifikation einer konkreten Person sei neben einer einzigartigen Kombination von GKV-Daten zusätzlich der Einsatz aktiver nachgehender Recherche erforderlich ("Einsatz eines Privatdetektives"). Dieses Szenario ist in der Forschungspraxis unrealistisch und sollte nicht länger als wichtiges Datenschutzrisiko herangezogen werden. Als Möglichkeiten der Missbrauchskontrolle wurde ein Loggen der Analysetätigkeit durch Aufzeichnung der Datenausgabe (Bildschirm) bei der Projektbearbeitung diskutiert. Alle Beteiligten sind sich darüber einig, dass angemessene Regelungen zur Vermeidung von Scientific Misconduct gelten und bei Missachtung entsprechend wirksame Maßnahmen folgen müssen.

Differential Privacy

Unter *differential privacy* werden verschiedene Verfahren zusammengefasst, die das Re-Identifizierungsrisiko in pseudonymisierten oder anonymisierten Datensätzen verringern sollen. Grundprinzip ist es, die Ausprägungen in einzelnen oder allen Variablen eines Datensatzes durch Addition oder Subtraktion eines Wertes zu verändern. Ziel ist es dabei, die Änderungen so vorzunehmen, dass die aus den veränderten Datensätzen gebildeten statistischen Größen sich möglichst wenig von den korrekten Werten abweichen, die sich aus den unverrauschten Daten ergeben [48].

Für einfache Statistiken wie Durchschnittswerte kann gezeigt werden, dass diese Abweichungen gering sind. Eine offene Frage ist jedoch, ob differential privacy Verfahren auch für die typischerweise multivariablen Modellierungen in Analysen der Versorgungsforschung anwendbar sind. Eine aktuelle Übersichtsarbeit kommt zu dem Schluss, dass sich das Verfahren für die Anwendung in der Gesundheitsforschung noch in einem frühen Entwicklungsstadium befindet und bislang wenig Anwendungsbeispielen aus der Praxis vorliegen [49]. Die Autor:innen schreiben: „*Significant gaps exist (...) for applications involving explanatory modeling and statistical inference, which are particularly important in epidemiology and clinical research. These typically involve estimating the effect of an exposure or intervention (eg, relative risk or odds ratio) on a particular health outcome, with statistical uncertainty measured by standard errors and confidence intervals. Statistical hypothesis tests are also critical tools for exploratory and confirmatory investigation (eg, such as to assess the efficacy of a novel health intervention). (...) Common examples might be follow-up time points in longitudinal trials, or clustered observations originating from the same individual or community unit. Information leakage with dependent data is one known weakness of standard definitions of differential privacy, so recent extensions have begun to address this. Future*

algorithm development efforts and assessments should take these into account.“ und an anderer Stelle: “(...) there are few differentially private algorithms available for inferential statistics commonly used in health research. Compromises to accuracy are also of concern in cases of limited data. Unfortunately, real-world evaluations of differential privacy in the health sector are extremely scant” (S. 2274). Mit den Anwendungsbeispielen Korrelationen, longitudinale Analysen, Zusammenführung von Versorgungsdaten, bspw. Daten aus mehreren Sektoren zu einem Patienten sprechen die Autorinnen und Autoren des aktuellen Reviews die typischen Nutzungen der Daten des FDZ in der Gesundheits- und Versorgungsforschung an. Die Nutzbarkeit von differential privacy ist für diese nicht geprüft.

Die verminderte Genauigkeit durch differential privacy ist speziell bei kleinen Datensätzen problematisch ist ([49], S. 2269). Dieser Einwand ist für die typischen Nutzungsszenarien der FDZ Daten von grundsätzlicher Bedeutung, da für relevante Aussagen für eine definierte Patientengruppe in aller Regel zahlreiche Variablen gleichzeitig berücksichtigt werden müssen. Dadurch entsteht eine hohe Anzahl von Merkmalskombinationen, von denen sehr viele nur noch begrenzte Fallzahlen enthalten. Da die verringerte Genauigkeit mehrere Variablen betrifft, hat sie eine unmittelbare Auswirkung auf eine Vielzahl von Merkmalskombinationen. Die Unsicherheiten infolge der Verrauschung potenzieren sich dadurch und die Interpretierbarkeit der Ergebnisse wird weiter eingeschränkt.

Eine in 2019 veröffentlichte Masterarbeit am Institut für Informatik der Freien Universität Berlin aus der Arbeitsgruppe ID Management untersuchte die Praktikabilität von *differential privacy* bei Datenauswertungen im Kontext von maschinellem Lernen („Maschine Learning“). Das Fazit bestätigt die Schlussfolgerungen von Ficek et al. [49]:

“We found that the adaptation of such a machine learning method to implement DP [Differential Privacy] is non-trivial and that even when applied carefully, privacy always comes at the price of accuracy. On our dataset, even the better-performing differentially private linear regression with a reasonable level of privacy produces a mean squared error twice as high as the normal linear regression on non-privatized data” [50].

Die Präzision der Ergebnisse leidet also erheblich unter der differential privacy. Studienergebnisse aus Daten des Zensus 2020 in den USA zeigen, dass die Anwendung von differential privacy zum Schutz der Privatsphäre zu einer erheblichen Verzerrung der COVID-19-Sterblichkeitsraten führt, was die Fähigkeit, die Pandemie zu verstehen, beeinträchtigt [51, 52]. Hauer et al. berichten: *“This distortion is particularly large for population groupings with fewer than 1,000 persons: 40 percent of all county-level age-sex groupings and 60 percent of race groupings.” ([51], S. 1). Hier besteht somit das Risiko einer Diskriminierung bestimmter Minderheiten innerhalb der Bevölkerung. Zu diesem Schluss kamen auch Santos-Lozada et al.: „Our results suggest that differential privacy will more strongly affect mortality rate estimates for non-Hispanic blacks and Hispanics than estimates for non-Hispanic whites. We also find*

significant changes in estimated mortality rates for less populous areas, with more pronounced changes when stratified by race/ethnicity.“ ([52], S. 13405).

Zusammenfassung und Schlussfolgerung: Die Anwendung von Differential Privacy im Rahmen des Forschungsdatenzentrums würde bei nachfolgender Verwendung der verrauschten Daten zu Forschungszwecken keine verlässlichen Ergebnisse liefern. Ganz im Gegenteil könnten aus Analysen auf der Basis von Differential Privacy wesentliche Fehlschlüsse resultieren, die ggf. erhebliche Auswirkungen auf die Gesundheitsversorgung in Deutschland haben können. Bei Anwendung von Differential Privacy kann das Ziel des Forschungsdatenzentrums, durch Nutzung der Daten zu Forschungszwecken die Gesundheitsversorgung der Patienten zu verbessern nicht mehr gewährleistet werden. So wären unter anderem Analysen zu Subpopulationen (beispielsweise geriatrische Patienten ab 70 Jahren) nicht mehr zuverlässig, weil aufgrund der durch das Verfahren in die Originaldaten eingebrachten zufälligen Fehler die geeignete Studienpopulation nicht mehr sicher zuordenbar wäre, die statistische Unsicherheit der Ergebnisse der Analysen sich erhöht und systematische Verzerrungen wahrscheinlicher werden.

Dezentrale Datenspeicherung und -auswertung

Eine dezentrale Datenspeicherung und -auswertung die Zuverlässigkeit der Ergebnisse typischer gesundheits- und versorgungsrelevanter Fragestellungen erheblich einschränken. Die soziodemographischen Charakteristika (Versicherten- und Morbiditätsstruktur) und daraus resultierend die Erkrankungshäufigkeiten und Inanspruchnahmen von Leistungen von Versicherten unterscheiden sich zwischen den am FDZ beteiligten z. Zt. N=97 Gesetzlichen Krankenkassen teilweise erheblich [53]. Systematischen Strukturen in den Ausgangsdaten (sogenannte Cluster oder Level) müssen bei der Wahl der Methoden zur statistischen Auswertung angemessen berücksichtigt werden.

Die Anwendung von Multilevel-Analysen (in der Literatur auch als mixed-effects models bezeichnet) ist eine übliche Technik zur Berücksichtigung solcher Cluster-Unterschiede [54]. Der Hintergrund solcher Analysen soll hier anhand eines einfachen Beispiels verdeutlicht werden: Wenn man beispielsweise die Wirkung eines neuen Rehabilitationskonzeptes auf die Dauer der Krankschreibung von Patienten nach einem Herzinfarkt untersuchen möchte, müssen (1) individuelle Faktoren auf der Ebene einzelner Patienten berücksichtigt werden, wie beispielsweise das Alter, kardiovaskuläre Vorerkrankungen. Weitere Einflussfaktoren (2) betreffen Gruppen von Patienten – diese bestehen auf der strukturellen und regionalen Ebene, z. B. die Verfügbarkeit des Rehakonzeptes, Wartezeiten, Indikationsstellung, Prüf- und Anerkennungspraxis durch die Krankenkassen. Die Auswirkungen des Rehakonzeptes auf die Dauer der Krankschreibung nach einem Herzinfarkt kann sich deshalb zwischen verschiedenen Krankenkassen unterscheiden. Dies gilt auch dann, wenn die individuellen Eigenschaften bei allen Versicherten gleich oder sehr ähnlich wirken. Werden diese beiden Ebenen nicht

gleichzeitig in die Analyse einbezogen, würden die Einflüsse auf Krankenkassenebene die Ergebnisse für die individuellen Faktoren verzerren.

Gruppenunterschiede (hier auf Ebene der Krankenversicherungen) werden in statistischen Analysen deshalb typischerweise mittels Multilevel-Analysen berücksichtigt. Für die Durchführung von Multilevel-Analysen mittels deterministischer Verfahren ist es zwingend notwendig, dass alle Daten in einem Datensatz vorliegen. Vergleicht man die Ergebnisse von separaten Regressionsanalysen mit den Werten aus einer mehrstufigen Regressionsanalyse (Multilevel-Analyse) an einem zusammengeführten Datensatz, so sind die Ergebnisse der separaten Analysen variabler, weisen also eine höhere Schwankungsbreite auf [54]. Entsprechend wären die Ergebnisse der statistischen Auswertung bei einer dezentralen Datenspeicherung und -auswertung weniger robust und würden bspw. zwischen den Krankenkassen unter Umständen unterschiedliche Ergebnisse liefern. So würde man bei einer dezentralen Auswertung des zuvor beschriebenen Beispiels bei den Versicherten einiger Kassen einen Effekt des Rehabilitationskonzeptes auf die Dauer der Krankschreibung erkennen, bei anderen nicht. Im Rahmen von Multilevel-Analysen kann dagegen der Einfluss des Rehakonzeptes um den Einfluss kassenbezogener Faktoren auf die Versicherten bereinigt werden. Nur so kann der tatsächliche Effekt des neuen Rehakonzeptes quantifiziert werden. Multilevel-Analysen werden in Analysen der Wirksamkeit, Sicherheit und Akzeptanz neuer Maßnahmen im Gesundheitswesen in aller Regel erforderlich sein.

Relevanz von Studien mit Record-Linkage allgemein

Unter den Forschenden besteht Einigkeit, dass die Zukunft der Analyse versorgungsnaher Daten in der Verknüpfung von GKV-Routinedaten mit anderen Daten liegen werde. Dies sei heute schon der Goldstandard (siehe NAKO, Innovationsfondsprojekte) und für die Beantwortung relevanter Versorgungsfragen sowie kausale Schlussfolgerungen in analytischen Projekten fast immer notwendig. Daher sei für den Erfolg des FDZ Gesundheit die Linkagemöglichkeit der Daten mit anderen Datenquellen eine entscheidende Voraussetzung. Nur so könne das FDZ Gesundheit den erhofften Impact auf die Forschungslandschaft haben. GKV-Routinedaten und epidemiologische Daten, Registerdaten oder primäre Studiendaten ergänzen sich, so dass durch Record-Linkage „das Beste aus beiden Welten“ zusammengebracht werden könne. Umweltbezogene Expositionsdaten, Daten zum Verhalten, Soziodemographie, Lebensqualität und viele weitere je nach Fragestellung wichtige Informationen fehlten in GKV-Routinedaten und müssten durch Record-Linkage mit anderen Quellen zusammengeführt werden, um zentrale wissenschaftliche Fragestellungen angemessen beantworten zu können. Die Verlinkung von Informationen aus der ePA wird ebenfalls als dringend erforderlich erachtet. Dadurch könne beispielsweise die Analyse von intersektoralen und regionalen Patientenpfaden deutlich verbessert werden.

Szenarien von Studien mit Record-Linkage über FDZ Gesundheit oder direkt über Krankenkassen

Es besteht Einigkeit, dass im Record-Linkage mit GKV-Routinedaten über das FDZ Gesundheit ein sehr großes Potential für die Verbesserung der Qualität, Sicherheit und Wirtschaftlichkeit der Patientenversorgung liegt, welches alle Akteure im Gesundheitssystem unbedingt nutzen müssen. Für die Umsetzung sei eine sehr gut konstruierte und effektive Datentreuhänderstelle von zentraler Bedeutung. Diese müsse strikt getrennt von den Versorgungsdaten operieren. Das fördere auch das Vertrauen in der Bevölkerung. Die NAKO zeige, dass die Akzeptanz der Bevölkerung zur Nutzung und Linkage von GKV-Routinedaten mit Primärforschungsdaten groß sei. Über 95% der NAKO-Teilnehmenden hätten in die Nutzung und das Linkage eingewilligt. Bei der NAKO werden auch PKV-Daten genutzt. Die Daten der PKV sollten perspektivisch unbedingt auch in das FDZ Gesundheit integriert werden.

Es ist nach übereinschätzender Einstimmung der Forschenden von entscheidender Bedeutung, dass ein Record-Linkage zukünftig über das FDZ Gesundheit möglich ist. Als primärer Schlüssel zwischen Primärdaten und den Daten des FDZ Gesundheit ist aktuell nur die Krankenversicherungsnummer (KVNR) möglich, da nur diese Angabe als identifizierendes Merkmal (in pseudonymisierter Form) vorliegt. Es gebe in Deutschland keine zentrale Stelle, die kassenübergreifend die KVNR mit Klartextangaben zu Name und Geburtsdatum vorhält. Die Datentreuhänderstelle am RKI könnte und sollte eine solche zentrale Rolle übernehmen. Diese KVNR liege aber in Studien nicht immer vor, so dass auch andere Möglichkeiten des Linkage notwendig sind. Hierfür bieten sich die identifizierenden Merkmale Name, Vorname, Geburtsdatum und Wohnadresse an, die bei der Datenübertragung von den KK an die Treuhandstelle übermittelt werden sollen. Mittelfristig wird von den Forschenden die Einführung eines *unique identifiers* gefordert, wie dies in anderen Ländern seit langem der Fall ist.

Es wird jedoch auch nach Inbetriebnahme des FDZ Gesundheit weiterhin Szenarien geben, in denen das Linkage mit GKV-Daten nur durch direkte Beteiligung von einzelnen Krankenkassen oder einer Gruppe von kooperierenden KK möglich ist. Aufgrund dieses Bedarfs sollten im Rahmen des Gesundheitsdatennutzungsgesetzes auch Regelungen zur forschungsfreundlichen Umsetzung von GKV-Datenanalysen durch Forschende in direkter Kooperation mit KK getroffen werden. Momentan liegt die Unterstützung von Linkage-Studien im Belieben der angefragten Krankenkassen. Jede Krankenkasse müsse einzeln angefragt werden und es sei häufig nicht transparent, warum ein Forschungsprojekt unterstützt wird oder nicht. Hier müssten dringend Regelungen oder Mechanismen gefunden werden, wie auf Basis der Einwilligung von Patientinnen und Patienten zum Datenlinkage dieses auch ohne „Klinkenputzen“ bei allen KK in Deutschland durch die Forschenden planbar umgesetzt werden könne. In diesem Zusammenhang sollte auch eine Klarstellung oder Neuregelung des §75 SGB X getroffen werden. Ziel ist es, dass dessen Anwendung bei Vorliegen einer

schriftlichen Einwilligung der Studienteilnehmenden entfällt und dadurch die Antragsstellung an die Aufsichtsbehörden der Krankenkassen vermieden wird.

Szenarien mit Record-Linkage von Registerdaten und Routinedaten

In diesem Gutachten stehen die NAPKON-Kohorten, namentlich die NAPKON-HAP beispielhaft für die Notwendigkeit eines Linkage mit den GKV-Daten aus dem FDZ. Diese Notwendigkeit besteht analog auch für viele weitere Kohorten, worauf hier grundsätzlich hingewiesen werden soll. NAPKON-HAP steht dabei für Kohorten, in denen eine individuelle informierte Einwilligung der Kohorten-Teilnehmende hinsichtlich des Datenlinkage mit GKV-Daten eingeholt wurde. Ein bereits etabliertes Beispiel ist die NAKO Gesundheitsstudie, ein vom BMBF, den Ländern und der Helmholtz-Gemeinschaft bereits seit mehr als 10 Jahren gefördertes Langzeit-Projekt einer bevölkerungsrepräsentativen Kohorte. Im Rahmen der NAKO werden zufällig aus der Wohnbevölkerung ausgewählte Personen in 18 über Deutschland verteilten Regionen zur Teilnahme eingeladen. Im Rahmen der Patienteninformation wird die Verknüpfung mit Kassendaten erläutert, mit denen sowohl Informationen über die Inanspruchnahme von Leistungen im Gesundheitssystem wie Krankenhausaufenthalte und Besuche bei aus- und Fachärzten erhoben werden sollen, als auch Komorbiditäten, Medikamente, spezielle Therapien, und weitere Diagnosen im prospektiven Verlauf. Die große Mehrheit der Probandinnen und Probanden erteilt ihre Einwilligung und auf dieser Basis wurden in den vergangenen Jahren bereits zahlreiche Verträge zur Datenübergabe mit einzelnen gesetzlichen und privaten Krankenkassen geschlossen. Die Datenanfrage erfolgt über die Treuhandstelle an die betreffende Krankenkasse, die Datenübergabe dann an das Integrationszentrum der NAKO. In Deutschland existiert eine Vielzahl weiterer Register mit Aufgaben in der medizinischen Forschung. In einem Gutachten im Auftrag des BMG wurden über 350 verschiedene Register zusammengestellt [11]. Diese sind in der Mehrheit Diagnose- oder Krankheitsgruppenbezogen und verfolgen häufig klinische und/oder versorgungsmedizinische Fragestellungen (z. B. Psoriasis-, Mukoviszidose, Herzinfarkt-, Asthma-, kindliche entzündliche Darmerkrankungen-Register). Eine weitere wichtige Gruppe von medizinischen Registern registriert Patienten, die bestimmte Therapien erhalten haben und verfolgt diese prospektiv, um deren Effektivität bezogen auf das therapeutische Ziel, aber auch Aspekte der Sicherheit, mögliche Risiken, und patientenbezogene Effekte wie Funktion und Lebensqualität zu messen (Aortenklappen-Register, Traumaregister, Implantateregister etc.). Für viele von diesen Registern ist ein Linkage mit Krankenkassendaten vorgesehen, für viele weitere medizinische Register wäre das sinnvoll zur Erreichung der Registerziele. Im Falle des Vorliegens einer informierten Einwilligung kann in der Vertrauensstelle im RKI ein individuelles Linkage erfolgen. Wenn in einem Register mit informierter Einwilligung die KVNR erhoben wird, erfolgt das Linkage anhand der KVNR. Wo diese nicht vorhanden ist, sollten die Kassendaten in der Vertrauensstelle anhand der identifizierenden Patienteninformationen (Name, Vorname, Geburtsdatum, Wohnadresse) zugeordnet werden, was jedoch bisher in der Vertrauensstelle

am RKI aktuell nicht möglich ist. Diese Möglichkeit der Datenlinkage auf der Basis identifizierender Angaben der Versicherten sollte deshalb perspektivisch vorgesehen werden.

Eine spezielle Gruppe sind die klinischen und epidemiologischen Krebsregister, die auf bundesgesetzlicher Basis von den Ländern betrieben werden (§ 65c SGB V, div. Landesgesetze). Da diese der onkologischen Qualitätssicherung dienen, besteht auf der Seite aller Behandelnden Meldepflicht. Auf der Seite der Betroffenen Patientinnen und Patienten besteht ein individuelles Widerrufsrecht. Das Widerrufsrecht beschränkt sich dabei auf die über die Datenerhebung der epidemiologischen Krebsregister hinausgehenden klinischen Informationen (gegen die epidemiologische Krebsregistrierung besteht kein Widerspruchsrecht). Die Krebsregister sind wichtige Datenquelle für Kohorten und andere klinische Register (z. B. Endpunkte-Validierung in der NAKO-Gesundheitsstudie, Untersuchung des Krebsrisikos unter den Teilnehmende klinischer diagnose- oder therapiebezogener Register). Zur Erfüllung des Zweckes der Krebsregister ist gleichzeitig aber auch eine Verknüpfung mit Kassendaten außerordentlich sinnvoll - da hier kein individuelles Einverständnis vorliegt, soll das Linkage nach den nachfolgend dargestellten Verfahren stattfinden. Diese Forderung wurde gleichsinnig von den befragten Forschenden wie auch von Vertretenden der Krankenkassen geäußert.

Auch in Kohorten, in denen keine individuelle informierte Einwilligung vorliegt, kann ein Linkage mit FDZ-Daten erforderlich sein. Als Beispiel für ein Datenlinkageprojekt eines Registers mit Routinedaten ohne individuelle informierte Einwilligung sei hier auf ein laufendes Innovationsfondsprojekt VersKiK [55] hingewiesen. In diesem werden die Therapiedaten des Deutschen Kinderkrebsregisters mit GKV-Routinedaten der Krankenkassen personenbezogen verlinkt, um unter anderem Spätfolgen als Resultat der Erkrankung selbst oder der initialen Therapie zu untersuchen. Als Kooperationspartner konnten zahlreiche Krankenkassen gewonnen werden, die zusammen einen Anteil von deutlich über 50% der GKV-Versichertenpopulation aufweisen. Besondere Herausforderung in diesem Projekt ist das Personenmatchen über chiffrierte Klartextangaben (Kontrollnummern und Bloomfilter). Die Funktionen Matching, Recordlinkage, Aufbereitung und Analyse der Daten erfolgen aus datenschutzrechtlichen Erwägungen und mit Umsetzung des Trennungsgebots in diesem Projekt räumlich, organisatorisch und rechtlich getrennt in mehreren (Vertrauens-)stellen und mit kaskadierender Pseudonymisierung. Datenschutzkonzept und die Datennutzung wurden von den zuständigen Aufsichtsbehörden der Krankenkassen genehmigt.

Die Kooperationsbereitschaft von Krankenkasse für dieses Projekt kann im Vergleich zu anderen vergleichbaren Projekten als sehr hoch angesehen werden. Ein Grund ist in dem Forschungsthema mit der Bezugspopulation von Kindern und Jugendlichen mit Karzinomerkrankung zu vermuten. Allerdings wäre auch hier eine vollständige Abbildung aller Krankenkassen aufgrund der relativ niedrigen Fallzahlen (im unteren 5-stelligen Bereich) als wünschenswert anzusehen. Dies könnte durch das Linkage mit Daten aus dem FDZ Gesundheit

umgesetzt werden, im Idealfall sogar Daten von sehr lang zurückliegenden Beobachtungszeiträumen. Bei den Krankenkassen dürfen die Abrechnungsdaten aufgrund von Löschrufen nur für eine bestimmte Zeit aufbewahrt werden. Daten aus dem ambulanten ärztlichen Bereich mussten bisher nach fünf Jahre gelöscht werden, durch Änderung der gesetzlichen Vorgabe ist dies auf 10 Jahre verlängert worden (§304 SGB V, Digitale-Versorgung-Gesetz (DVG) vom 09.12.2019). Das FDZ Gesundheit hingegen muss versichertenbezogene Einzeldatensätze erst nach 30 Jahren löschen (§303d (3) SGB V). Allerdings ist hier zu beachten, dass diese Daten sukzessive von den Krankenkassen jahresweise geliefert werden. Damit steht ein längerer Beobachtungszeitraum erst nach entsprechender Vorlaufzeit zur Verfügung.

Zudem muss in Zusammenhang mit dem Datenbestand von Registern darauf hingewiesen werden, dass die zur Verlinkung mit den Daten des FDZ notwendige Krankenversicherungsnummer (KVNR) nur ausnahmsweise (beispielsweise in den Klinischen Krebsregistern nach § 65c SGB V) erhoben wurde. Auch im oben beschriebenen Linkageprojekt mit den Daten des Kinderkrebsregisters fehlte diese Angabe. Das Linkage musste daher auf Basis von personenidentifizierenden Daten erfolgen. Zukünftig muss auch im FDZ Gesundheit die Voraussetzung geschaffen werden, eine Verlinkung auf der Basis der identifizierenden Daten (Name, Vorname, Geburtsdatum, Wohnanschrift) zu ermöglichen. Dazu ist es notwendig, dass diese Daten in der Vertrauensstelle am RKI gehalten werden (siehe hierzu Kapitel 4.1.2 Identitätsmanagement).

3.3 Sicht der Ethik

Im Rahmen der Gutachtenerstellung führten die Gutachter eine Diskussion zu ethischen und rechtlichen Aspekten der wissenschaftlichen Nutzung von GKV-Routinedaten mit besonderem Fokus auf das konsentbasierte Linkage mit anderen Daten. Wir danken den teilnehmenden Expertinnen und Experten: Prof. Dr. Eva Winkler, Prof. Dr. Sabine Salloch, Prof. Dr. Kai Cornelius und Prof. Dr. Jan Schildmann.

Nach Umriss des Gutachtenauftrags wurden insgesamt 11 Fragen beantwortet, die 3 Fragenblöcken zugeordnet waren. Einige Schlüsselfragen waren den Expertinnen und Experten vorab übermittelt worden. Im Laufe des Gesprächs ergaben sich teilweise weitere Fragen und Gesichtspunkte, die dann direkt aufgegriffen wurden. Im Folgenden sind die Hinweise und Empfehlungen dieser nach den Frageblöcken zusammen gestellt. Es wurde kein formelles Konsenverfahren angewandt, sondern offen unter Moderation von Gutachtern (Prof. Schmitt, Dr. Ihle) diskutiert. Zu den meisten Punkten herrschte unter den Expertinnen und Experten Einigkeit. Wo dies nicht der Fall war, ist dies in den folgenden Ausführungen explizit angegeben.

Fragenblock 1: Linkage über (zukünftige) Infrastruktur FDZ Gesundheit

Frage 1: *Wie sollte die Infrastruktur zur Arbeit mit den Daten des FDZ Gesundheit gestaltet sein?*

Empfehlungen:

o Entscheidend ist ein angemessenes Verhältnis von Sicherheit der Infrastruktur und Praktikabilität der Umsetzung der Forschung.

o Absolut sichere Umgebung ist nicht realistisch erreichbar bzw. würde Forschung stark behindern, wenn nicht gänzlich verhindern - deshalb sollte eine vernünftige Risikominimierung das Ziel sein.

o Infrastruktur sollte daher die Forschungsinstitute einbeziehen. Wenn diese noch zu definierende Anforderungen erfüllen, sollte die Arbeit mit den Daten des FDZ Gesundheit und gelinkten Forschungsdaten an den Forschungsinstituten stattfinden können.

o Bei vorliegender informierter Einwilligung (Studienspezifische Einwilligung sowie Broad Consent zum Linkage mit GKV-Routinedaten) halten die Ethiker nach einheitlicher Auffassung die Datennutzung für ethisch unbedenklich und befürworten die Herausgabe der Daten an geeignete Forschungseinrichtungen (studienpezifischer Export aus dem FDZ)

o Wenn Probanden zum Linkage mit „Krankenkassendaten“ oder „GKV-Routinedaten“ einwilligen, so beinhaltet diese Einwilligung die Nutzung der genannten Daten über die Krankenkasse sowohl direkt als auch über das FDZ Gesundheit

o Für das Linkage im vorliegenden Use Case ist einerseits die Einwilligung (NAPKON-Erhebung + Linkage) und andererseits eine gesetzliche Grundlage für die GKV-Daten (nach § 303a-f SGB V) gegeben. Damit sind die normativen Anforderungen für das Linkage der zwei forschungsethisch verschiedenartigen Datenquellen erfüllt ("Gute Grundlage")

Frage 2: *Welche Anforderungen sind an die Vertrauensstelle/Datentreuhänderstelle zu stellen? Welche Daten sollten dort vorliegen?*

Empfehlungen:

o Auch Kohorten ohne Information zur KVNR müssen auf Basis eines Informed Consent mit GKV-Daten verknüpfbar sein. Hierfür ist das Vorhalten von Identitätsdaten in einer unabhängigen Vertrauensstelle notwendig, um bspw. für Probanden in Studien die KVNR nachzuerfassen

o Datenschutzrechtlich sind Identitätsdaten wie Namen/Adressen der KVNR gleichgestellt. Auch in NAKO/NAPKON liegt das Wissen über Identität und KVNR bereits vor, damit kommt kein Zusatzwissen ins System und die individuelle Verknüpfung ist damit ethisch unproblematisch und führt nicht zu einer Änderung der Risikobewertung

o Da keine doppelte Pseudonymisierung in der Vertrauensstelle vorgesehen ist, stellt sowohl aus ethischer als auch aus datenschutzrechtlicher Sicht die Speicherung von Name, Geburtsdatum und Adresse in der Vertrauensstelle zusätzlich zur KVNR (die auch als Identitäts-Angabe gilt) hohe Anforderungen an die Datensicherheit der Vertrauensstelle.

o Das Wissen über Identitäten ist sowohl auf Seiten von NAPKON als auch GKV gegeben, daher ergibt die Kombination kein Zusatzwissen

o die hohe Zahl der Kassenwechsler ist relevant. Diese könnten aber bei einer Doppelpseudonymisierung nicht gemanaged werden, daher ist die originale KVNR notwendig, bzw. ein Pseudonym basierend auf der KVNR.

o Rolle der Vertrauensstelle muss klar definiert sein (technische Aufgabe oder Forschungsakteur) → wichtige Zusatzaufgabe wäre eine Umsetzung von Widerruf, aber in der aktuellen Konstellation nicht das Kontaktieren von Patienten

o Der Begriff "Vertrauensstelle" oder "Treuhandstelle" sollte noch einmal überdacht werden, da beide normativ sehr aufgeladen sind. Laien leiten aus der Bezeichnung möglicherweise einen Anspruch ab, der hier nicht komplett erfüllt wird (unterstellter Laienanspruch: "Ich vertraue darauf, dass die Vertrauensstelle dafür sorgt, dass meine Daten in meinem Sinne genutzt werden")

o Bislang hat die Vertrauensstelle nur technische Aufgaben und sollte dann nicht Vertrauensstelle heißen, da damit mehr Erwartungshaltung verbunden ist, die man auch nicht mehr einlösen kann

o In der Infrastruktur des FDZ Gesundheit nimmt die Vertrauensstelle eine neutrale Rolle ein, daher wäre ein neutraler Begriff evtl. passender.

o Andererseits hat die Vertrauensstelle durchaus die Aufgabe, „vertrauensvoll“ mit den Daten umzugehen. Auch das sichere Verwahren einer Identitätsnummer ist eine vertrauensvolle Aufgabe, was wiederum für die Beibehaltung der Bezeichnung spricht

o Bei der Abwägung der Bezeichnungen "Treuhandstelle" und „Vertrauensstelle“ wird letztere bevorzugt, weil „Treuhandstelle“ eine "Stellvertreterfunktion" suggeriert, die nicht gegeben ist.

Frage 3: *Wie sehen Sie die Rolle der Ethikkommission bei dem beschriebenen Szenario eines Linkage von Kohortendaten mit Daten des FDZ Gesundheit?*

Empfehlungen:

o Beratung durch Ethikkommission wird grundsätzlich empfohlen à Einschätzung „Probandenschutz“ sowie Forschungsprojekt selbst (Sozialer Wert, methodische Qualität, bedenkliche/diskriminierende Forschung)

o Bei Nutzung von Daten des FDZ Gesundheit ohne informierte Einwilligung und ohne Linkage ist Beratung durch EK nicht zwingend erforderlich. Ob für ein beantragtes Auswertungsprojekt ein Ethik-Votum eingeholt werden sollte, könnte bspw. durch das Use&Access Board festgelegt werden (Klarstellung in Gesetz/Verordnung wünschenswert) [Thema wurde nicht tiefer besprochen]

o Beim beschriebenen Szenario (Linkage mit Einwilligung) ist Beratung durch EK notwendig.

Frage 4: *Halten Sie im beschriebenen Szenario einen Antrag nach §75 SGB X für erforderlich?*

Empfehlungen

o Bei Vorliegen einer spezifischen informierten Einwilligung scheint aus ethischer Sicht ein §75-Antrag verzichtbar, da alle anderen Aspekte der §75-Prüfung bereits durch andere Gremien geprüft werden (Ethikkommission, ggf. Use&Access Board)

o Beim Broad Consent wird hingegen wegen der fehlenden Spezifik der Einwilligung eine zusätzliche Governance als Sicherungsmaßnahme notwendig. Diese kann einerseits durch den §75-Antrag als gewährleistet angesehen werden. Andererseits ist auch ein Use&Access Board eine weitere Sicherungsmaßnahme und wird daher für das FDZ Gesundheit empfohlen (dann wäre ein §75-Antrag auch bei Broad Consent nicht zwingend erforderlich).

Frage 5: Wie ist ein Reidentifizierungspotenzial durch Datenherausgabe aus dem FDZ Gesundheit zu bewerten?

Empfehlungen

o Forschungsinstitute, die bestimmte Standards erfüllen, sollten als Teil der Infrastruktur zur Datennutzung (und nicht als Außenstehender) verstanden werden (s. Empfehlungen Frage 1). Dies kann z. B. durch den Abschluss von verbindlichen Verträgen zwischen dem FDZ und den Institutionen der Datennutzer:innen unterstützt werden.

o Es ist wichtig, nicht durch restriktive Regeln einen Hemmschuh gegen die Forschung aufzubauen

o Eine Datenherausgabe aus dem FDZ ist ethisch und rechtlich nicht anders zu bewerten, als bisherige Datennutzungen über einzelne Kassen mit Datenherausgabe

o Es existieren praktikable, erprobte Lösungen bei den Krebsregistern mit vertraglichen Möglichkeiten z. B. code of conduct

o Außerdem ist in §303e auch ein Reidentifizierungsverbot enthalten!

o Normatives Misstrauen ist grundsätzlich hinderlich

o Es wäre besser, klare Kriterien und Regeln vertraglich festzuschreiben bzw. in einem Code of Conduct festzuhalten. Die Regeln bzw. der Code of Conduct könnte Teil eines ohnehin angedachten Qualifizierungs-Curriculums für FDZ-Datennutzer sein

o Ein Mitloggen der vorgenommenen Auswerteprozesse wäre als zusätzliche Sicherungsmaßnahme denkbar, erscheint aber nicht zwingend notwendig

o Zu unterscheiden ist die unbeabsichtigte Vernachlässigung von Sicherheitsmaßnahmen von einem bewussten Missbrauch für kriminelle Zwecke. Ersteres ist durch Schulung /Anforderungen der Nutzenden zu minimieren (z. B. Zertifikat in Anlehnung an GCP-Zertifikat). Letzterer Fall ist in Deutschland bisher nicht bekannt und müsste juristisch geahndet werden.

Fragen 6: Welche Zeiträume der Daten (vor- und rückwirkend) sind für GKV-Daten angemessen bei entsprechender Einwilligung?

Empfehlungen

o Wichtiger als der absolute Zeitraum (3 oder 5 Jahre oder lebenslang) ist die Möglichkeit, jederzeit und niederschwellig den Widerruf der Einwilligung in das Linkage zu ermöglichen (z. B. wenn man nach 3 Jahren eine Krankheit bekommt, von der man nicht möchte, dass der Forscher diese sieht)

o Für welchen Daten-Zeitraum die Einwilligung gegeben wird, kann auch individuell durch die Patientinnen und Patienten festgelegt werden: Elegant wäre hier ein flexibler Datenzeitraum, den der Patient bestimmen kann, z. B. durch Auswahlmöglichkeit in der NAPKON-POP-Einwilligung

o Zu Bedenken ist, welche Forschung nicht möglich ist, wenn man den Zeitraum eingrenzt

Fragenblock 2: Linkage über Krankenkassen

Frage 7: *Erwächst für Forschende durch die Einwilligung von Personen zum Linkage ihrer GKV-Daten für Forschungszwecke ein Anspruch, diese Daten tatsächlich von den betreffenden Krankenkassen zu erhalten?*

Kommentare/Empfehlungen:

- o Es ist wichtig, hier im Blick zu haben, welche Gründe eine mögliche Zurückhaltung von Kassen haben könnte: z. B. zusätzlicher Aufwand oder eigene Forschungs- (oder auch Geschäfts-) interessen
- o Kassen haben gegebenenfalls auch Interessenskonflikte
- o Die Zurückhaltung kann zu einem Publikationsbias führen
- o Ethisch gilt, dass Kassen ihren Versicherten verpflichtet sind: Beitragszahlungen der Versicherten, Solidarfinanzierung des gesamten Systems
- o Damit sollte die Datennutzung der Kassendaten als Nutzen für das Gemeinwohl gesehen werden. Sofern "ein hohes Gemeinwohlnutzenpotential" besteht und noch dazu mit Auftrag des Versicherten per Einwilligung die Datenübermittlung gewünscht ist, bestehen gute Gründe, den Willen und Auftrag des Probanden /Versicherten umzusetzen, da dieser ja die Kontrolle über seine Daten haben muss.
- o Sofern der Zusatzaufwand für die Kassen finanziert/honoriert wird, müsste die Argumentation pro Datennutzung ausfallen
- o Für Transparenz und Fairness sollte eine Struktur eingerichtet werden, die darstellt, warum und wie Entscheidungen für oder gegen ein Projekt fallen
- o Gegebenenfalls könnte sich die Problematik durch das FDZ künftig auflösen, aber eventuell erst nach einigen Jahren, daher muss die direkte Anfrage durch Forschende bei den Kassen aktuell noch mitbetrachtet werden
- o Es ist durch eine entsprechende Governance sicherzustellen, dass Datennutzungen planbar werden (sowohl über Einzel-Kassen als auch FDZ)
- o Eine "Strafbewehrung" der Nichtumsetzung von Datennutzungen durch Kassen/FDZ würde entsprechende gesetzliche Grundlagen erfordern
- o Forschungsprojekte mit direktem Kontakt zu Patientinnen und Patienten, z. B. per Einladungsschreiben, müssen weiter möglich sein (auch langfristig bei vollständig funktionalem FDZ Gesundheit), sonst wäre das forschungsethisch kritisch. Daher sollte durch das Gesundheitsdatennutzungsgesetz auch die Möglichkeit der wissenschaftlichen Nutzung und des Linkage von GKV-Daten über die Krankenkassen geregelt werden.

Frage 8: *Erwächst für Studienteilnehmende ein Anspruch, dass ihre Daten den Forschenden durch die KK übergeben werden?*

Kommentare/Empfehlungen:

- o Wenn einzelne Versicherte einen derartigen Anspruch geltend machen würden, bestünde eine große Selektionsgefahr
- o Zudem ist für das vorliegende Szenario zumeist kein Eigennutzen für Versicherte gegeben
- o Damit steht eher eine faire Auswahl der Studienteilnehmer im Vordergrund und das Argument wäre schwer zu führen.

Frage 9: *Wie ist aus forschungsethischer Sicht die Rolle der PKV-Daten einzuschätzen?*

Kommentare/Empfehlungen:

- o Besondere Datenschutzrisiken wären mit der Nutzung von PKV-Daten nicht gegeben
- o Möglicherweise müsste die Unsicherheit (auch Einbeziehung der PKV) gesetzgeberisch angegangen und ein Mandat für die Forschung definiert werden (Manche Privatversicherungen machen in NAKO mit, in die Krebsregisterfinanzierung sind über ihren Verband alle PKV eingebunden)
- o Für PKV-Versicherte erwächst kein direkter Nachteil, dass ihre Daten nicht ins FDZ Gesundheit eingehen, weil in der Regel kein direkter individueller Nutzen aus der Forschung zu erwarten ist (anders als bei Interventionsstudien).

Fragenblock 3: Generelle Fragen zur Sammlung und Nutzung von GKV-Daten im FDZ Gesundheit

Frage 10: *Welche Hinweise / Empfehlungen haben Sie zur momentanen administrativen Prüfung von Anträgen zur Nutzung von GKV-Routinedaten?*

Kommentare/Empfehlungen:

- o Zu §75-Antrag siehe oben.
- o Datennutzungen werden mitunter mehrfach redundant geprüft (mehrere EKs & Datenschutzprüfung durch DSB und EK)
- o Das pragmatische Konzept der "führenden EK" existiert (erste EK prüft, andere schließen sich dem Votum an) und könnte den Prüfprozess beschleunigen und verschlanken
- o Doppelprüfungen sind Ausdruck der Unsicherheit von Zuständigkeiten
- o Gegebenenfalls besteht hier gesetzlicher Regelungsbedarf

Frage 11: *Welche Hinweise / Empfehlungen haben Sie Frage des Widerspruchsrechts bei der Nutzung von GKV-Routinedaten?*

Kommentare/Empfehlungen:

- o Widerspruchsrecht ist Teil der informationellen Selbstbestimmung, weshalb Ethiker generell eher für Widerspruchslösung votieren
- o Bei Abwägungen ist Qualität der Forschung - insbesondere das Potenzial von Bias - zu berücksichtigen
- o Wenn man den Nutzen aus der Forschung gut erklärt, erhöht sich auch die Einwilligungsquote. Man kann jedoch nicht stillschweigend davon ausgehen, dass jeder einverstanden ist.
- o Aus ethischer Sicht wäre eine Abwägung von Nutzen der Forschung und Bias aufgrund von Selektion bei Nicht-Einwilligung entscheidend.
- o Eine Lösung ohne Widerspruchsrecht erfordert einen sehr hohen Nutzen aus der Forschung bzw. starke Limitationen bei Widerspruch einzelner Personen/Personengruppen. Eine Verweigerung des Widerspruchs wäre nur in eng umgrenzter Nutzung möglich
- o Votum der Ethiker ist daher grundsätzlich für differenzierte Opt-out-Lösung anstatt genereller Verweigerung des Widerspruchs
- o Bei Opt-in Lösung wäre der erwartete Bias erheblich, so dass Opt-in in diesem Kontext nicht bevorzugt wird (viele Forschungsfragen könnte man dann gar nicht untersuchen)

3.4 Sicht der Patientenvertretung

Um Meinungen und Ansichten von Patientenvertretenden in die Fragestellung des Gutachtens einzubeziehen, wurden Ansprechpersonen von insgesamt 14 Patientenvertretungs- und Selbsthilfegruppen eingeladen, an einer schriftlichen Befragung teilzunehmen, darunter die Deutsche Rheuma-Liga, Selbsthilfegruppe Long COVID Deutschland, Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung, chronischer Erkrankung und ihren Angehörigen (BAG SELBSTHILFE), die Deutsche Leukämie- & Lymphom-Hilfe, die Patienten-Hilfe der Deutschen Sarkom-Stiftung, der Bundesverband "Das frühgeborene Kind", der Sprecher des Koordinierungsausschusses der Patientenvertretung im GBA, der NCT-Patientenbeirat, das Haus der Krebshilfe und der Patientenbeirat des Deutschen Zentrums für Neurodegenerative Erkrankungen (DZNE).

Die jeweiligen Ansprechpersonen wurden in einem persönlichen Anschreiben über den Kontext der Befragung informiert (Gutachtenauftrag, Einrichtung des FDZ Gesundheit, Datenlinkage) und erhielten einen Auszug aus der Patienteninformation der NAPKON-POP Kohortenstudie (Modul GKV-Routinedaten). Das Anschreiben und der Fragebogen finden sich in der Anlage 3. Anhand des Fragebogens wurde der Grad der Zustimmung oder Ablehnung zu 12 Fragen auf einer 5-stufigen Likertskala erhoben (siehe Anlage 3).

Insgesamt gingen Antworten von 10 Patientenvertretenden ein.

Alle 10 Patientenvertretenden befürworteten es, dass auf Basis der informierten Einwilligung durch Patientinnen und Patienten (Informed Consent) NAPKON-Studiendaten mit Krankenkassendaten individuell verknüpft und für die Gesundheitsforschung genutzt werden (7/10 Patientenvertretenden gaben "starke Zustimmung", 3/10 "eher Zustimmung"). 9/10 Teilnehmende würden den Mitgliedern Ihrer Patientenorganisation zur Studienteilnahme raten, ein Teilnehmer hatte hierzu "keine klare Meinung".

Jeweils 4/10 Befragte stimmten "sehr stark" bzw. "eher" der Aussage zu, dass ein Auftrag an die Politik besteht, die Nutzung von Krankenkassendaten und deren Verbindung zu Studiendaten (z. B. NAPKON) durch entsprechende Rahmenbedingungen zu ermöglichen. 2/10 Befragten lehnten diese Aussage "eher ab". Alle Patientenvertretenden stimmten darin überein, dass Krankenkassen beauftragt sind, ihre Daten für die Forschung verfügbar zu machen, wenn die Versicherten ausdrücklich eingewilligt haben (je 5/10 Patientenvertretenden gaben "starke Zustimmung" und "eher Zustimmung"). Forschende sollen bei vorliegender Einwilligung zur Datenverknüpfung diese auch tatsächlich durchführen können (4/10 „starke Zustimmung“, 5/10 "eher Zustimmung", 1/10 "keine klare Meinung"). Nach mehrheitlicher Auffassung (8/10) der Teilnehmenden erwächst durch die Einwilligung in die Datennutzung für die teilnehmenden Patientinnen und Patienten ein (moralischer) Anspruch, dass die vorgesehene Datenverknüpfung auch tatsächlich realisiert wird.

Ein großer Vorteil der Krankenkassendaten ist es nach Ansicht der Patientenvertretende für die Wissenschaft, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung der Patientinnen und Patienten einen Langzeitverlauf abbilden können, sofern die Forschungsfragen diesen Zeitraum notwendig machen. Auf die Frage *„Welcher zukünftige Zeitraum sollte durch eine gegebene Einwilligung abgedeckt werden bzw. ist aus Ihrer Sicht hier angemessen und für Patientinnen und Patienten absehbar?“* antworteten 2/10 „lebenslang“, 2/10 "10 Jahre", 3/10 „5 Jahre“ und "3/10" Teilnehmende gaben an, dass dies variabel, in Abhängigkeit der Fragestellung, sowie bis zum Widerspruch angemessen ist. Damit würden 7/10 Patientenvertretende prospektive Laufzeiten von mehr als 5 Jahren unterstützen, wenn dies zur Beantwortung einer sinnvollen Forschungsfrage nötig ist.

9/10 der teilnehmenden Patientenvertretende stimmten dafür, dass auch Private Krankenkassen die Versorgungsdaten ihrer Versicherten z. B. im skizzierten Beispiel für die Forschung verfügbar machen sollten. Ein Teilnehmer hatte hierzu keine klare Meinung. Alle 10 Patientenvertretende sahen in der Tatsache, dass privat Krankenversicherte Ihre Daten nicht in gleichem Maße für die Forschung verfügbar machen können wie gesetzlich Krankenversicherte, einen Nachteil für die PKV-Versicherten. Der Forderung, dass nach Abschluss

einer Studie die Ergebnisse interessierten Studienteilnehmenden zur Verfügung gestellt werden sollten, gaben 9/10 Teilnehmerinnen „starke Zustimmung“.

Am Ende des Fragebogens hatten die teilnehmenden Patientenvertretenden noch die Möglichkeit, weitere Anregungen und Kommentare zu äußern. Ein Patientenvertreter forderte von der Politik *„alles dafür zu tun, um den momentan geltenden Datenschutz in Deutschland zu entschärfen, damit die Forschung gerade im Bereich der Krebsforschung wie in anderen europäischen Ländern zu ermöglichen und dadurch auch die internationale Verknüpfung von Daten zu ermöglichen.“* Und führte weiter aus, dass *„Insbesondere Krebspatienten, aber auch Hinterbliebene ein sehr großes Interesse haben, ihre persönlichen Daten der Forschung zur Verfügung zu stellen – meine Erfahrung aus vielen Patientengesprächen.“*

3.5 Sicht der Klinikerinnen und Kliniker (FOSA)

Befragung des NUM-FOSA-Fachbeirats

Über eine anonyme Online-Umfrage, die über die im Netzwerk Universitätsmedizin (NUM) eingerichteten fächer- und organspezifischen Arbeitsgruppen (FOSAs) durchgeführt wurde, sollten die Meinungen und Standpunkte der Kliniker und klinischen Wissenschaftler in die Empfehlungen des Gutachtens einfließen. Zuvor wurden in einer Sitzung des NUM-FOSA-Beirats, in dem die Sprecher aller FOSA organisiert sind, der Gutachtenauftrag sowie die in der Umsetzung befindliche Machbarkeitsstudie (siehe Kapitel 1.3) zur Verknüpfung von NAPKON-Daten mit Daten der gesetzlichen Krankenversicherungen vorgestellt.

Der verwendete Fragebogen ist im Anhang zu finden (siehe Anhang). Der Fragebogen enthielt Fragen zur Person der "Befragten", Fragen zu bestehenden Erfahrungen mit der Nutzung von GKV-Daten, Fragen zur Datenverknüpfung, Fragen zu Barrieren bei der Datennutzung und zu Hilfsmöglichkeiten. Insgesamt wurden 14 Fragen gestellt, von denen 7 Freitextfragen waren.

Insgesamt wurden von 30 Befragten Fragebogen ausgefüllt. Etwa die Hälfte der 29 bestehenden NUM FOSAs waren durch Vertretende (14/29) abgedeckt. Es lagen 28 auswertbare Fragebögen vor.

Ausnahmslos wurde von allen Antwortenden (n=26) befürwortet, dass auf der Basis einer informierten Einwilligung durch Patientinnen und Patienten externe Daten, zum Beispiel GKV Routinedaten von Krankenkassen, Registerdaten und Befragungsdaten, individuell verknüpft und für die Gesundheitsforschung genutzt werden.

Die Antworten auf die Frage: „Welche konkreten Forschungsfrage(n) halten Sie für besonders geeignet um sie mittels Verknüpfung mit GKV-Routinedaten von Krankenkassen im Kontext

von NAPKON zu beforschen?“ spiegeln das breite Forschungsspektrum wider in dem mit GKV-Daten gearbeitet werden sollte.

Neben den bereits klassischen Forschungsinhalten von GKV-Routinedatennutzung (Arbeitsunfähigkeit, Inanspruchnahmemuster und deren Prädiktoren, Krankheitskostenermittlungen, Untersuchung von Versorgungspfaden und Leitlinienadhärenz) wurden insbesondere auch Datenverknüpfungsanwendungen genannt (z. B. Langzeitverlaufsuntersuchungen von Post-COVID Symptomen verknüpft mit AU-Zeiten, direkten und indirekten Krankheitskosten im Vergleich zu Propensity Score gematchten Kontrollen ohne COVID-19, Identifizierung von früheren Erkrankungen als Risikofaktoren für Post-COVID Symptome). Weitere Vorschläge waren vergleichende Längsschnittuntersuchungen mit gesunden Kohorten, risikofaktoren-basierte Fragestellungen, Mortalitätsstatistiken, die Untersuchung von seltenen Erkrankungen und Medikamenten-Interaktionen. Ein weiterer Punkt waren auch Studien zu Endpunkten in registerbasierten RCTs. Die vollständige Liste aller Freitextantworten befindet im Anhang.

Etwa die Hälfte der Befragungsteilnehmenden hat bereits Erfahrung mit Abrechnungsdaten der Krankenkassen, meisten im Bereich von 2-5 Projekten. Ein kleinerer Teil der Befragungsteilnehmenden hat auch bereits Verlinkungen von primären Daten mit anderen Datenquellen durchgeführt - hier meist epidemiologische Studien und (Krebs-)registerstudien (z. B. NAKO, SHIP, KORA und weitere).

Das Potenzial von Sozialdaten wird von der Mehrheit der Befragten als sehr hoch oder hoch eingeschätzt (23/25). Das Potenzial der GKV-Routinedaten zeigt sich unter anderem in dem breiten Indikationsfeld, das die befragten Wissenschaftler nach eigenen Angaben bereits abdecken: Neben Infektionskrankheiten (v.a. Harnwegsinfektionen), Arzneimittelforschung ("Langzeitverwendung von Opioiden in Deutschland inkl. Mortalität und Risikofaktoren für Missbrauch"), Krankheitskosten (z. B. Fibromyalgie-Syndrom und Reizdarmsyndrom), wurden Fragen aus dem Bereich chronischer Erkrankungen ("Gebrechlichkeit", "Mehrdimensionales, gesundes Altern und biologisches Alter", Psychiatrie und onkologische Versorgungsforschung) bearbeitet. Entsprechend breit sind auch die Forschungsbereiche /Forschungsfragen, die die NUM Forschenden für besonders relevant halten, um sie zukünftig mittels GKV-Routinedaten zu beforschen: Neben allgemeineren Fragen zur Erforschung ganzer Indikationen ("COVID-19") oder bestimmter Sektoren des Gesundheitssystems (z. B. "Rehabilitation" - Einsatz von Heilmitteln und Einfluss auf Aktivität und Teilhabe/Arbeitsfähigkeit, insbesondere unter Berücksichtigung extrabudgetärer Diagnosen - Langzeitrehabilitation - wo werden Patienten wie versorgt? Welche Lücken gibt es? Rehabilitation im Sozialraum - wo fehlen Ansprechpartner im System?)" wurde Potenzial für Kostenstudien und Arzneimittelstudien identifiziert. Geeignet seien die GKV Routinedaten auch, um intersektorale und regional vergleichende Fragestellungen zu bearbeiten und die Wirkung von unterschiedlicher politisch-ökonomischer

Rahmenbedingungen (z. B. Vergütungsstrukturen, Personalanforderungen) auf die Qualität und Kosten der Versorgung zu untersuchen.

Mit Hilfe von fünfstufigen Likert-Skalen wurde nach der Wichtigkeit/Relevanz verschiedener Barrieren für die Initiierung neuer Projekte mit Krankenversicherungsdaten gefragt und welche dieser Hindernisse als am relevantesten gegen die Initiierung neuer Projekte angesehen wurden.

Für die meisten der genannten Barrieren wurde deren Bedeutung als sehr relevant oder relevant eingeordnet. Als besonders hinderlich wurden „Schwierigkeiten mit dem Zugang und der Kontaktaufnahme zu Krankenkassen“, „Fehlende Kooperationsbereitschaft/ mangelnde Ressourcen von Krankenkassen“ und „Fehlendes eigenes technisches Know-how / datenwissenschaftliche Kompetenzen zur Verarbeitung der GKV Daten“ eingestuft (siehe Tabelle 4).

Tabelle 4: Barrieren für die Initiierung neuer Projekte mit Krankenversicherungsdaten

	Bereits selbst erlebt? (Ja/Nein)	Relevanz (Nennungen)				
		Sehr relevant (1)	(2)	(3)	(4)	Irrelevant (5)
Schwierigkeiten mit Zugang und Kontaktaufnahme zu Krankenkassen	15/9	11	7	3	3	0
Fehlende Kooperationsbereitschaft/ mangelnde Ressourcen von Krankenkassen	14/10	4	11	8	1	0
Schwierigkeiten bei der Beantragung der Beratung durch Ethikkommissionen	10/14	4	4	5	7	4
Ablehnung bzw. hohe Auflage bei der Beratung durch Ethikkommissionen	5/19	1	3	10	6	4
Ablehnung bzw. Auflage bei der Bewertung durch Datenschutzbeauftragte	11/13	8	7	6	1	2
Ablehnung bzw. Auflage bei der Bewertung von Anträgen nach § 75 bei den zuständigen Landesbehörden	4/20	2	6	12	3	1
Fehlendes eigenes technisches Know-how / datenwissenschaftliche Kompetenzen zur Verarbeitung der GKV Daten	-	5	7	9	2	2

Weitere Hinweise auf Hürden/Barrieren bei der Nutzung von GKV-Routinedaten und insbesondere bei deren Verknüpfung mit Individualdaten waren Kosten, lange Wartezeiten bis zum Datenfluss, der retrospektive Charakter der Daten, die Verzögerungszeit von der Erhebung bis zur Übermittlung und die Nachverfolgung von Patienten bei einem Kassenwechsel. Fehlende methodische Beratung bei der Planung, Durchführung und Auswertung von Studien, die Einbindung vieler regulatorischer Akteure in multizentrische Studien sowie individuelle Interessen der beteiligten Krankenkassen und deren je nach kooperierender Krankenkasse unterschiedlichen Konzepte der Zusammenarbeit wurden ebenfalls als Hürden identifiziert. Auch der unzureichende Zugang zu (validen) PKV Daten wurde als Problem identifiziert.

Aus wissenschaftlicher Sicht ist ein großer Vorteil von Krankenkassendaten, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung des Patienten einen Langzeitverlauf abbilden können, wenn Forschungsfragen diesen Zeitraum erforderlich machen.

Unter der Prämisse, dass die Einwilligung jederzeit widerrufen werden kann, stellt sich die Frage, für welchen zukünftigen Zeitraum der Datenerhebung ab der erteilten Einwilligung die Patienten ihre Zustimmung geben könnten. Der kürzeste von den Wissenschaftlern angegebene angemessene Zeitraum war 5 Jahre (n=8), aber die Mehrheit der Wissenschaftler hielt einen lebenslangen Zeitraum für angemessen (n= 12). Zweimal wurden 10 Jahre als angemessener Zeitraum genannt.

Tabelle 5: Unterstützungsbedarfe

	Relevanz (Nennungen)				
	Sehr relevant (1)	(2)	(3)	(4)	Irrelevant (5)
Zentrale Aufbereitung und Bereitstellung von GKV Routinedaten	15	6	0	0	0
Zentrale Qualitätssicherung der Daten	15	5	1	0	0
Einheitliches zentrales Verfahren für regulatorischen Anforderungen (Ethik, Datenschutz, §75 Genehmigungsverfahren)	16	4	1	0	0
Bereichsübergreifende Regelung zur Einwilligung der Patienten zur Verarbeitung der Behandlungsdaten, sofern Patienten zustimmen (z. B. Broad Consent)	18	3	0	0	0

Zentrale Unterstützungsangebote (Bedarfe) wie die Aufbereitung und Bereitstellung von GKV Routinedaten sowie eine zentrale Qualitätssicherung der Daten wurde ebenso für sehr relevant gehalten, wie eine bereichsübergreifende Regelung zur Einwilligung der Patienten zur Verarbeitung der Behandlungsdaten (z. B. Broad Consent). Hohe bis sehr hohe Relevanz als Unterstützungsleistung hätte laut der befragten Wissenschaftlerinnen und Wissenschaftler auch ein einheitliches zentrales Verfahren für regulatorischen Anforderungen (Ethik, Datenschutz, §75 Genehmigungsverfahren).

Darüber hinaus wurde die Gelegenheit genutzt, weitere Maßnahmen zu benennen, die die Nutzung der GKV-Routinedaten (und deren Verknüpfung) erleichtern könnten. Neben der Stärkung der interdisziplinären Zusammenarbeit zur Erhöhung der Qualität der Ergebnisse, wurde von den Wissenschaftlerinnen und Wissenschaftlern angeregt, zu klären, welche Zugriffsrechte existierten und wem die Daten (nicht) gehörten. Vorgeschlagen wurde die Einrichtung einer Vertrauensstelle sowie der Austausch mit Ländern, die bereits Erfahrungen mit Routinedaten haben. Auch die Verwendung eines einheitlichen Datenformats wurde vorgeschlagen.

In den abschließenden allgemeinen Hinweisen und Anmerkungen wurde der Wunsch geäußert, dass eine Verknüpfung mit Daten der Rentenversicherungsträger insbesondere für den Bereich der Rehabilitation zusätzlich wünschenswert wäre. Die Befragten meldeten zurück, dass die Durchführungsmodalitäten für das FDZ von konkreter Bedeutung seien. Die Zentralisierung sollte nicht mit hohen Verwaltungskosten und zu langen Wartezeiten einhergehen. Die Kosten sollten überschaubar bleiben.

4 Empfehlungen

4.1 Empfehlungen zum FDZ Gesundheit

4.1.1 Allgemeine Empfehlungen zum FDZ Gesundheit

Die Potenziale des FDZ Gesundheit werden von Forschenden, Krankenkassenvertretenden und nicht zuletzt von den Gutachtern als sehr hoch angesehen. Mit dem FDZ Gesundheit besteht die große Chance für eine breit angelegte, evidenzgeleitete Verbesserung der Patientensicherheit, der Qualität, der Effektivität und der Effizienz der Gesundheitsversorgung und damit der Gesundheit der Bevölkerung der Bundesrepublik auf der Basis von Forschung mit GKV-Routinedaten. Um mit Hilfe des FDZ Gesundheit die wissenschaftliche Nutzung von GKV-Routinedaten zu fördern und das Potenzial der Forschung mit GKV-Routinedaten zu heben, muss das FDZ Gesundheit forschungsfreundliche Nutzungsmöglichkeiten etablieren. Es ist daher entscheidend, praktikable Instrumente und Verfahren für die Forschenden zur Nutzung der FDZ-Daten aufzubauen, sowie für begleitende Gremien und Strukturen (z. B. Use&Access Board) Rahmenbedingungen zu schaffen, die ein effektives Arbeiten unterstützen. Wichtige Charakteristika einer forschungsfreundlichen Nutzungsmöglichkeit beinhalten eine einfache, unbürokratische Antragstellung, eine schnelle Datenbereitstellung nach Antragsgenehmigung und einen möglichst geringen Zeitverzug zwischen Datenentstehung und Datenbereitstellung im FDZ. Nach Prüfung der Voraussetzungen ist ein direkter Datenzugriff durch die Forschenden wichtig. Wichtig ist zudem die Evaluation der Nutzung und Nutzungsmöglichkeiten des FDZ Gesundheit innerhalb eines relativ kurzen Zeitraums nach Inbetriebnahme (z. B. nach 2 Jahren), um eventuelle Optimierungsbedarfe frühzeitig erkennen und Optimierungen umsetzen zu können. Hierzu ist eine Transparenz der Prozesse im FDZ Gesundheit erforderlich, beispielsweise in Form regelmäßiger (z. B. jährlicher) Berichte über die bearbeiteten Nutzungsanträge inklusive der dabei benötigten Fristen, aufgetretener Probleme und erkannter Barrieren.

Die Datennutzung kann für Daten des FDZ Gesundheit ohne individuelles Linkage mit anderen Daten i. d. R. per Remote-Nutzung erfolgen. Im Fall des individuellen Linkage ist eine Übergabe der verknüpften Daten vom FDZ Gesundheit an Forschungsinstitute eine bedeutsame Variante der Datennutzung (konkrete Szenarien werden nachfolgend aufgeführt). Eine Übergabe von GKV-Daten an Forschungsinstitute ist seit langem gängige Praxis bei Forschungsprojekten, welche mit Beteiligung einzelner Krankenkassen realisiert werden. Somit entsteht durch eine Datenherausgabe aus dem FDZ Gesundheit keine grundlegend neue Situation, welche komplett neu zu bewerten wäre. Die Forschungsinstitute der Datennutzer sollten für eine Datenübergabe definierte infrastrukturelle Voraussetzungen erfüllen. Dies sind insbesondere die Erfüllung der Anforderungen hinsichtlich Datenschutz- und Datensicherheit sowie der Fähigkeit zur qualifizierten Aufnahme und Haltung der übergebenen Datenstruktur. Im Einklang mit den Empfehlungen der befragten Ethikerinnen und Ethiker empfehlen wir daher, dass Forschungsinstitute, die bestimmte Voraussetzungen

(auditierbar) erfüllen, als Teil der Infrastruktur zur Nutzung der Daten des FDZ Gesundheit angesehen werden und die Analysen der (gelinkten) Daten des FDZ Gesundheit in der Regel an den Forschungsinstituten stattfinden. Hierzu empfehlen wir vertragliche Vereinbarungen zwischen dem FDZ und den Institutionen der Antragsteller. Die zu den Studienteilnehmenden auf Basis einer informierten Einwilligung individuell verlinkbaren FDZ-Daten sollen normalerweise aus dem FDZ-Datensatz extrahiert und exportiert werden. Erforderlich ist dies jedoch auch bei probabilistischem Linkage der GKV-Daten mit Daten der Klinischen Krebsregister, wie dies z. B. im Rahmen der WiZen-Studie [26, 33] erfolgreich umgesetzt wurde.

Anders als etwa in der Klinischen Forschung ist es bei GKV-Routinedatenanalysen nicht möglich, alle Einzelheiten der Datenaufbereitung und -analyse vor der Sichtung des Datenkörpers im Detail bereits festzulegen. Vielmehr müssen häufig vorab festgelegte Regeln (z. B. im Sinne eines Studienprotokolls/-plans) während der Projektbearbeitung im Rahmen des Analyseprozesses datenbasiert spezifiziert oder angepasst werden, z. B. wenn weitere Plausibilisierungen, Sensitivitätsanalysen oder neue Operationalisierungen von Variablen notwendig werden. Daher ist ein breiter Zugriff auf die Datenbasis am FDZ entscheidend, der auch Queranalysen an Daten/Populationen ermöglicht, die nicht zwangsläufig direkt mit dem beantragten Forschungsthema in Beziehung stehen müssen. Zusätzlich sollte die Möglichkeit geschaffen werden, einen breiten Datensatz für Forschende zu exportieren. Beide Zugangswege, breiter Zugriff auf die FDZ Daten und Export eines breiten Forschungsdatensatzes, sind hierbei nicht als sich gegenseitig ausschließende Alternativen zu sehen, sondern sollten in Forschungsprojekten auch parallel beschritten werden können. Zur Erstellung der Auswerteskripte ist eine Bereitstellung eines Teildatensatzes i. d. R. ausreichend. Simulierte Daten sind dagegen nicht geeignet, weil durch die angewandten Syntheseverfahren wichtige Zusammenhänge im Realdatenkörper nicht abgebildet sein können, sodass die darauf erstellte Syntax nicht auf den Realdaten lauffähig ist oder/und aufeinander aufbauende Analyseschritte nicht sinnvoll konzipiert werden können.

Einige für die Gesundheitsversorgung äußerst relevante Fragestellungen können nur mit einem vollständigen, krankenkassenübergreifenden Datensatz aller Versicherter in der notwendigen wissenschaftlichen Qualität bearbeitet werden. Hierzu zählen beispielsweise Fragestellungen der Qualitätssicherung, der Patientensicherheitsforschung, der Arzneimittelsicherheit, des Versorgungs(qualitäts-)monitorings, zu seltenen Erkrankungen sowie zur Leitliniengerechtigkeit der Versorgung in spezifischen Patientengruppen. Die Einrichtung des FDZ Gesundheit ermöglicht die Forschung in diesen wichtigen Arbeitsfeldern. Aus diesem Grund sollte die Übermittlung der GKV-Daten an das FDZ Gesundheit ohne Widerspruchsmöglichkeit der Versicherten erfolgen.

Der zukünftige Datenkranz des FDZ Gesundheit (ab Beobachtungsjahr 2019) stellt eine deutliche Erweiterung der bisher vom DIMDI bereitgestellten Morbi-RSA-Daten dar. Insbesondere die verfügbaren Leistungsdaten aus dem ambulanten (EBM-Angaben) und stationären Sektor (OPS-Angaben) bzw. die zusätzlichen Detailangaben (Inanspruchnahme nach Leistungserbringern/ Fachärzte und tagesgenaue Liegezeiten) bilden das Versorgungsgeschehen besser ab und ermöglichen damit eine bessere Bearbeitung vieler Forschungsfragen. Kritisch muss allerdings darauf hingewiesen werden, dass auch im FDZ-Datenumfang weitere Sektoren aus dem Bereich der Gesetzlichen Krankenversicherung nicht enthalten sind (z. B. Pflege, siehe hierzu Kap. 2.5)

Ein weiterer Nachteil besteht in der Tatsache, dass bisher keine Daten der Privaten Krankenversicherung (PKV) für das FDZ Gesundheit vorgesehen sind. Da sich Privatversicherte bzgl. des sozioökonomischen Status und damit auch des Gesundheits- und Inanspruchnahmeverhaltens tendenziell von gesetzlich Krankenversicherten unterscheiden, können je nach Fragestellung Selektionsprobleme und eine eingeschränkte Generalisierbarkeit der Ergebnisse resultieren [53, 56, 57]. Patientenvertretende sahen in der Tatsache, dass PKV-Versicherte ihre Daten nicht in gleichem Maße der Forschung zur Verfügung stellen können wie GKV-Versicherte, eine Benachteiligung und plädieren ebenso wie die Gutachter dafür, die PKV-Daten perspektivisch auch im FDZ Gesundheit verfügbar zu machen [58]. Nach Kenntnis der Gutachterkreises gibt es hierzu aus der PKV Bestrebungen, die Daten harmonisiert und strukturiert, ähnlich wie in der Gesetzlichen Krankenversicherung, aufzubereiten und auswertbar zu machen. Eine hohe Qualität soll danach die Aufbereitung der Daten des stationären Sektors haben, ähnliches wird für den Sektor Arzneimittel aus dem ambulanten Bereich berichtet. Auf dem AGENS Methodenworkshop 2023 in Köln fand zu diesem Thema ein Preconference-Workshop "Special Interest Group PKV-Daten" statt [59]. Auf dieser Veranstaltung wurde unter anderem darauf hingewiesen, dass auch die Beteiligung von Privaten Krankenversicherungen an Forschungsprojekten freiwillig ist und es hierdurch zu Selektionseffekten im Hinblick auf die einbezogene Versichertenpopulation kommen kann.

Aktuell ist die Datenlieferung von den Krankenkassen an das FDZ im jährlichen Turnus festgeschrieben. Je nach Forschungsinhalt oder Dringlichkeit von datenbasierten Entscheidungen (Pandemische Gefährdungslage, [12]) sollten Lieferungen im kürzeren Zeitabständen geprüft werden. Nachdem die technischen Prozesse implementiert und validiert sind, sollte eine sektorenbezogene Anpassung der Liefertermine erfolgen. Stationäre Aufenthalte werden den Krankenkassen wenige Tage nach Entlassdatum übermittelt, Verordnungsdaten sind innerhalb eines Monats verfügbar. Zudem sollte auch die quartalsmäßige Abrechnung der ambulanten ärztlichen Daten über die KVen an die Krankenkassen unter dem Gesichtspunkt der schnelleren Verfügbarkeit und Nutzung der Daten auf den Prüfstand gestellt werden. So könnten in einer Art "fast lane" Verfahren kritische Ereignisse tagesgenau berichtet werden, beispielsweise der Anstieg von Atemwegserkrankungen in der ambulanten Versorgung als Frühwarnsystem, noch bevor es zu einer Häufung von stationären Aufnahmen kommt. Dieses

Verfahren sollte flexibel auf die jeweilige (heterogene) Gefährdungslage reagieren können, da insbesondere der ambulante Sektor des Gesundheitswesens in kritischen Situationen häufig erste Anlaufstelle für die Patienten ist, dieser aber aktuell abrechnungstechnisch (Quartalsbezug) erst mit einem Zeitversatz von mehreren Monaten einer Nutzung in den Krankenkassen bzw. im FDZ Gesundheit verfügbar ist. Die vergangenen Jahre haben überdeutlich gezeigt, dass Deutschland im Vergleich zu unseren europäischen Partnerländern mit sehr deutlich größerem Zeitverzug und auch nur unvollständig auf Basis von regionalen oder selektierten Daten die Effekte der SARS-CoV-2 Pandemie auf die Gesundheitsversorgung ermitteln konnte. Während beispielsweise Holland nur wenige Monate nach Pandemiebeginn deren Effekte auf die Versorgung von Brustkrebserkrankten [60, 61] ermitteln und entsprechend gegensteuern konnte, ist in Deutschland die Evidenzlage zur Versorgung von Krebspatienten seit Beginn der Pandemie und der Einfluss auf die Prognose noch immer weitgehend unklar, weil eine entsprechende Datengrundlage fehlt. Mit dem FDZ Gesundheit muss dieser Standortnachteil Deutschlands dringend korrigiert und eine möglichst unverzügliche Datenverfügbarkeit in Krisensituationen sichergestellt werden.

Wie bereits beschrieben ist ein Versicherter/eine Versicherte im FDZ Gesundheit durch das so genannte Periodenübergreifende Pseudonym (PüP) als eindeutiger Identifikator gekennzeichnet. Das Periodenübergreifende Pseudonym wird in der Vertrauensstelle DaTraV (RKI) aus der KVNR durch ein Verfahren erzeugt, das bereits im Informationssystem Versorgungsdaten beim DIMDI, der Vorgängereinrichtung des FDZ Gesundheit, etabliert wurde. Durch Beibehaltung des Verfahrens ist gewährleistet, dass die bisherigen Daten des Informationssystem Versorgungsdaten mit den an das FDZ Gesundheit neu übermittelten Daten identische Pseudonyme erhalten und die Daten damit über die Zeit personenbezogen verlinkbar sind.

Für die Generierung des Periodenübergreifenden Pseudonyms aus der KVNR müssen in der Vertrauensstelle DaTraV (RKI) die KVNR der Versicherten vorliegen. Die hierfür notwendigen Prozesse sind in den Dokumenten "Pseudonymisierungskonzept zur Datenübertragung im Rahmen der DaTraV (Lieferpseudonym)" (Version 1.8) sowie "Pseudonymisierungskonzept zur Datenübertragung im Rahmen der DaTraV und für die Funktion Datenfreigabe Forschungsdatenzentrum der ePA (Periodenübergreifendes Pseudonym)" (Version 1.3) beschrieben und wurden in Meetings von Mitarbeitenden der Vertrauensstelle DaTraV erläutert [62, 63]. Die Dokumente wurden auf Bitten des BMG dem Gutachterkreis von der Vertrauensstelle DaTraV zur Verfügung gestellt. Grundsätzlich gilt hierbei, dass "im Rahmen dieses Verfahrens ... das RKI lediglich die Kombination von Lieferpseudonym und Arbeitsnummer" erhält. Die Krankenkassen generieren hierfür aus der KVNR einen Hashwert, der durch einen von der Vertrauensstelle DaTraV jährlich wechselnden Schlüssel (ein so

genannter Salzwert) gesalzen¹ wird. Die Rückrechnung von den Hashwerten auf die originale KVNR in der Vertrauensstelle des RKI erfolgt durch einen so genannten Brute-Force-Angriff bei bekanntem Schlüssel (Salzwert). Dabei werden in der Vertrauensstelle aus den möglichen Ausprägungen der KVNR (10-stellig, führender Buchstabe, 8 Ziffern und eine Prüfziffer) unter Kenntnis des Salzwertes die Hashwerte berechnet und somit eine Referenztafel erzeugt, mit Hilfe derer die übermittelten Hashwerte in die KVNR rückübersetzt werden.

Aus kryptologischer und technischer Sicht ist dieses Verfahren grundsätzlich anwendbar. Hashverfahren als technisch nicht rückrechenbare Falltürfunktion werden für so genannte Einwegverschlüsselungen eingesetzt, also genau in solchen Fällen, bei denen das (technische) Rückrechnen auf den Klartext wirksam verhindert werden soll. Darüber hinaus ist bei Anwendung eines Hashverfahrens die Kollisionsfreiheit zu prüfen, da bei Hashverfahren (technisch bedingt) aus zwei unterschiedlichen Ausgangswerten derselbe Hashwert erzeugt werden kann. Sollte dieser (sehr seltene) Fall unerkannt und ungeprüft eintreten, so würden einer Person eine falsche KVNR bzw. ein daraus abgeleitetes falsches Pseudonym (PüP) zugeordnet werden. Da der Salzwert durch die Vertrauensstelle DaTraV an die Krankenkassen (jährlich wechselnd) verteilt wird, könnte die Kollisionsfreiheit eines Schlüssels auf Basis aller möglichen Krankenversichertennummern vorab in der Vertrauensstelle geprüft werden. Dies sollte auch erfolgen, damit nur kollisionsfreie Salzwerte verwendet werden.

Alternativ könnte für die Übermittlung der KVNR an die Vertrauensstelle ein dafür explizit benanntes Verschlüsselungsverfahren angewendet werden, z. B. AES als symmetrisches Verfahren wie in [64] vorgeschlagen. Das Verfahren lässt sich technisch leicht implementieren und ist bei Wahl eines hinreichend langen Schlüssels (beispielsweise 256 Bit) als sicher anzusehen. Eine Prüfung auf Kollisionsfreiheit, wie bei dem jetzt eingesetzten Hashverfahren wäre nicht erforderlich. Der Einsatz eines asymmetrischen Verschlüsselungsverfahrens hingegen verbietet sich, da mit dem öffentlichen Schlüssel grundsätzlich Probeverschlüsselungen auf die endliche Menge an KVNR möglich sind. Die KVNR-Listen könnten aber auch über eine sichere Verbindung von den einzelnen Krankenkassen direkt an die Vertrauensstelle ohne Durchleitung im GKV-Spitzenverbands DaTraV übermittelt werden., beispielsweise mit Hilfe von Cryptshare ([Cryptshare - security made easy](#)), das unter anderem von den AOKen und der TK zur sicheren Datenübermittlung eingesetzt wird.

Strukturelle/funktionelle Empfehlungen

Stellvertretend für Studien, in denen Patienten rekrutiert werden und per Informiertem Konsent ihre Einwilligung in die Teilnahme an der Studie geben, wird hier von der NAPKON

¹ Das Salzen entspricht technisch der Konkatenation (Verkettung) des Originalwertes (hier: KVNR) mit dem Salzwert als Schlüssel zu einer einzigen Zeichenkette und anschließender Bildung des Hashwerts aus dieser Zeichenkette.

POP Kohortenstudie ausgegangen. Ziel ist es, die (klinischen) Primärdaten aus dieser NAPKON-Studie mit den Routinedaten im FDZ Gesundheit zu verlinken und zu analysieren.

Für die weitere Beschreibung und unsere entsprechenden Empfehlungen soll davon ausgegangen werden, dass in den jeweiligen Stellen nachfolgend aufgeführten Dokumente, Angaben bzw. Daten vorliegen. Eine zusätzliche juristische Prüfung der jeweiligen Rechtsgrundlagen zur Datenverarbeitung insbesondere für die hier erweiterte Funktionalität der Vertrauensstelle im RKI erfolgte im Rahmen dieses Gutachtens nicht. Grundsätzlich müssen für eine solche Erweiterung der Funktionalität alle relevanten Zuständigkeitsbereiche, einschließlich des BfDI, einbezogen werden.

In der NAPKON Studienstelle:

1. Einwilligungserklärung der Patientin/des Patienten inkl. Zustimmung zur Verlinkung der Primärdaten mit den Daten der Krankenkassen bzw. des FDZ Gesundheit
2. Identifizierende Klartextangaben der Patientin/des Patienten wie Vor- und Nachname, Geburtsdatum, Wohnadresse (an allen Standorten von NAPKON) und zusätzlich am Standort Würzburg die Krankenversicherungsnummer sowie die zum Zeitpunkt der Einwilligung versichernde Krankenkasse. Die Sammlung von KVNR und Kassenzugehörigkeit wird an allen Standorten vervollständigt.
3. Medizinische Primärdaten der Patientin/des Patienten

In der Vertrauensstelle des FDZ Gesundheit (RKI):

1. Algorithmus zur Erzeugung des Periodenübergreifenden Pseudonyms (PüP) aus der Krankenversicherungsnummer
2. hierfür notwendiger Schlüssel
3. Krankenversicherungsnummer
4. identifizierende Klartextangaben zum Patienten (Name, Vorname, Geburtsdatum, Wohnadresse)

Da sowohl KVNR als auch identifizierende Klartextangaben gegenwärtig nicht in der RKI-Vertrauensstelle vorhanden sind, ist die Nacherfassung der KVNR in NAPKON zwingend. Wo dieses nicht gelingt, kann der Patient beim Linkage nicht berücksichtigt werden. Dies führt unweigerlich zu einem Survivor-Bias, da Personen, die an Covid oder einer anderen Erkrankung verstorben sind, systematisch ausgeschlossen werden.

Im FDZ Gesundheit:

1. Periodenübergreifendes Pseudonym des/der Versicherten
2. Routinedaten aus der Gesetzlichen Krankenversicherung

Matching (Identitätsabgleich)

Für das Datenlinkage bzw. Recordlinkage zweier Quellen (hier NAPKON-Studiendaten und GKV-Routinedaten) bedarf es einer oder mehrerer gemeinsamer Schlüsselvariablen für das Matching. Hierbei müssen im Allgemeinen drei Fälle betrachtet werden:

- Matching auf Basis der Krankenversicherungsnummer
- Nacherhebung der Krankenversicherungsnummer aus Klartextangaben der Proband:innen
- Matching auf der Basis von identifizierenden Patientenvariablen

Diese Fälle werden im Folgenden ausgeführt.

Matching auf Basis der Krankenversicherungsnummer

Diese Variante stellt den einfacheren Fall dar. Wenn im Rahmen der Rekrutierung der Kohorte die Krankenversicherungsnummer erfasst wurde, so kann diese als Schlüsselvariable für das Matching direkt genutzt werden, da in der Vertrauensstelle des FDZ Gesundheit regelhaft daraus das periodenübergreifende Pseudonym (PüP) generiert werden kann. Das heißt, in beiden Datenquellen liegt dieselbe Schlüsselvariable vor und kann mit Hilfe des festgelegten Algorithmus in das PüP überführt werden. Anhand des PüP kann anschließend das Matching ausgeführt werden.

Nacherhebung der Krankenversicherungsnummer aus Klartextangaben der Probandinnen und Probanden

Wird hingegen die Krankenversicherungsnummer im Rahmen der Rekrutierung in einer Kohorte nicht erfasst, muss diese nachträglich erhoben werden. Hierfür stehen nach aktuellem Stand prinzipiell mehrere Möglichkeiten zur Verfügung:

- Erhebung bei der Probandin/beim Probanden
- Erhebung aus den Abrechnungsdaten der stationären Einrichtung
- Erhebung bei der Krankenkasse, bei der die Probandin/der Proband versichert ist
- Erhebung bei der Rentenversicherung BUND
- Erhebung in der RKI-Vertrauensstelle

Die *Erhebung beim Probanden* ist dann möglich, wenn im Rahmen der Einwilligung die Kontaktdaten erfasst wurden und der Patient einer Kontaktierung zugestimmt hat. Beides ist bei den NAPKON-Kohorten üblicherweise der Fall. Bei Vorliegen dieser Voraussetzungen kann die Patientin/der Patient postalisch oder telefonisch um Übermittlung der KVNR gebeten werden. Hierbei kann erfahrungsgemäß nicht von einer 100%igen Rücklaufquote

ausgegangen werden. Diese nachträgliche Erfassung der KVNR beim Probanden/bei der Probandin wird regulär in NAPKON POP in den Standorten Berlin und Kiel durchgeführt.

Die *Erhebung aus den Abrechnungsdaten der stationären Einrichtung* erfolgt in der stationären Einrichtung, in der die Patientin/der Patient versorgt wurde und in der die Rekrutierung stattgefunden hat. Bei GKV-Versicherten ist die KVNR als Ordnungskriterium im Rahmen der Abrechnung mit der versichernden Krankenkasse notwendigerweise dokumentiert. Der Abgleich ist technisch möglich, da die Klartextangaben der Patientin/des Patienten in der Einwilligung und im Krankenhausinformationssystem vorliegen. Zusätzlich kann der Zeitpunkt der Einwilligung mit den Liegezeiten innerhalb der stationären Einrichtung abgeglichen werden. Dieser Abgleich kann manuell oder (halb-)automatisiert erfolgen. Eine nahezu 100% Nacherfassung der KVNR sollte damit möglich sein. In wieweit dieser Abgleich rechtlich zulässig ist, muss jeweils geprüft werden.

Die *Erhebung bei der Krankenkasse* ist grundsätzlich möglich. Problematisch ist dies allerdings, wenn die versichernde Krankenkasse in einer Kohorte nicht erhoben wurde und damit alle Krankenkassen (zur Zeit n=97) einzeln kontaktiert werden müssten, und dort dann auf Basis der Klartextangaben der Patientin/des Patienten der Abgleich mit dem Mitgliederverzeichnis der jeweiligen Krankenkasse erfolgen müsste. Eine vollständige Nacherfassung ist hier nicht zu erwarten, die rechtliche Grundlage für dieses Vorgehen ist gegenwärtig nicht klar.

Die *Erhebung bei der Deutschen Rentenversicherung* scheint zum aktuellen Zeitpunkt logistisch einfacher als die Nachfrage bei allen Krankenkassen zu sein, also ein Abgleich der Klartextangaben der Probandin/des Probanden mit dem Verzeichnis der Rentenversicherung BUND in der Datenstelle der Rentenversicherung (DSRV). In diesem Verzeichnis sind alle Versicherten mit ihren identifizierenden Angaben enthalten, für die jemals eine Sozialversicherungsnummer oder synonym eine Rentenversichertennummer vergeben wurde. Da andererseits die Rentenversichertennummer in der Vertrauensstelle Krankenversichertennummer (VST KVNR) als Basis für die Generierung der Krankenversichertennummer genutzt wird, liegt für jeden GKV-Versicherten mit einer KVNR auch eine RVNR bei der Deutschen Rentenversicherung vor. Logistisch würde ein Abgleich folgendermaßen erfolgen:

- 1.) Anfrage durch die Studiengruppe NAPKON bei der Rentenversicherung Bund mit Hilfe der vorliegenden Klartextangaben der Probandin/des Probanden und Rückübermittlung der RVNR.
- 2.) In einem zweiten Schritt kann mit Hilfe der RVNR bei der VST KVNR die zugehörige KVNR erfragt werden. Dies ist ein Standardprozess, der zwischen Krankenkassen und Vertrauensstelle implementiert ist. Damit liegt der Studienstelle die KVNR des Probanden/der Probandin vor.

Die jeweiligen Prozesse sind bereits implementiert. Eine rechtliche Grundlage für die Nutzung dieser Prozesse für die hier beschriebene Nacherfassung der KVNR fehlt nach Ansicht der Gutachter allerdings noch. Diese könnte bspw. in einer Neufassung des § 75 SGB X ergänzt werden.

Matching auf der Basis von identifizierenden Patientenvariablen

Eine Ermittlung der KVNR für das Linkage mit den FDZ-Daten wäre in der Vertrauensstelle DaTraV am RKI möglich, wenn in der Vertrauensstelle die identifizierenden Klartextangaben aller Versicherten zusammen mit der KVNR vorlägen. Die aktuelle Aufgabenbeschreibung der FDZ-Vertrauensstelle sieht das Vorhalten einer solchen Referenzliste jedoch derzeit noch nicht vor. Die Prozesse in der Vertrauensstelle sehen z. Zt. vor, dass nach Erzeugung der Periodenübergreifenden Pseudonyme (PüP) und Übermittlung dieser Liste zusammen mit den Lieferpseudonymen (LP) an das FDZ alle in der Vertrauensstelle generierten temporären Tabellen wieder gelöscht werden. Um eine Erhebung der KVNR in der FDZ-Vertrauensstelle zu ermöglichen, sind hierfür eine Neufassung der rechtlichen Rahmenbedingungen sowie eine Anpassung der Prozesse mit Übermittlung von identifizierenden Klartextangaben der Versicherten zusammen mit der KVNR von den Krankenkassen an die FDZ-Vertrauensstelle erforderlich. Wenn zudem die RVNR, die den Krankenkassen regelhaft vorliegt und im System der Krankenkassen abgespeichert ist, an die Vertrauensstelle des FDZ Gesundheit im RKI übermittelt wird, wäre ein Abgleich der Klartextangaben mit dem Mitgliederverzeichnis der Rentenversicherung (siehe oben) möglich und auch eine Verknüpfung von KVNR basierten Datenbeständen (FDZ Gesundheit) und RVNR basierten Datenbeständen (beispielsweise des FDZ Rentenversicherung) ohne weiteres möglich.

Wie oben ausgeführt existieren relevante Forschungsbedarfe, bei denen auf der externen Seite (Kohorte, Register, klinische Studie) die KVNR nicht vorliegt. Um zu diesen Kohortenpatientinnen und -patienten, die ihre informierte Einwilligung hierzu explizit erteilt haben, die GKV-Daten zu matchen, muss ebenfalls der Abgleich auf der Ebene der individuellen identifizierenden Angaben erfolgen können (Name, Geburtsdatum, Adresse). Wir empfehlen daher, diese Angaben in der Vertrauensstelle DaTraV beim RKI vorzuhalten und dort prospektiv zu pflegen. Bislang werden in der Vertrauensstelle DaTraV nach Generierung der periodenübergreifenden Pseudonyme und Übermittlung an das FDZ Gesundheit alle temporären Dateien gelöscht. Bei Vorhalten der personenidentifizierenden Angaben müssen daher geeignete Verfahren des Datenschutzes und der Datensicherheit implementiert werden. Hierfür existieren bereits entsprechende Konzepte, beispielsweise in der NAKO-Vertrauensstelle oder der Datenstelle der Rentenversicherung, die als Blaupause dienen können.

Wenn die RKI-Vertrauensstelle zukünftig wie von den Gutachtern empfohlen die beschriebenen Klartextangaben von Versicherten zusammen mit der KVNR vorhält, könnte

diese perspektivische weitere Aufgaben in einer späteren Ausbaustufe der Rechtsgrundlage des FDZ, wie beispielsweise das Anschreiben von Versicherten, übernehmen (siehe hierzu Kapitel 4.1.2). Diese Empfehlung der Gutachter ist im Einklang mit der Argumentation der Ethiker, die an die Vertrauensstelle den Anspruch stellen, vertrauensvoll mit den Studienteilnehmenden und Probandinnen und Probanden zu kommunizieren. Auch der Widerspruch sollte ggf. von Studienteilnehmenden an die RKI-Vertrauensstelle kommuniziert werden.

Nach den bisherigen Ausführungen soll die Vertrauensstelle DaTraV (RKI) zukünftig zwei Funktionen übernehmen: Identitätsmanagement (Verwaltung der versicherten-identifizierenden Daten inkl. der KVNR) und Pseudonymisierungsdienst (Umschlüsselung der KVNR in das Periodenübergreifende Pseudonym (PüP)). Der Pseudonymisierungsdienst entspricht dabei in seinem Aufgabenprofil der aktuellen Vertrauensstelle DaTraV. Das neu zu implementierende Identitätsmanagement könnte hierbei in eine eigenständige Stelle, ähnlich der Treuhandstelle in NAPKON oder der NAKO, ausgelagert werden. Die räumliche und organisatorische Trennung beider Aufgaben unterstreicht, dass die Zuordnung des Periodenübergreifenden Pseudonyms (nach Umschlüsselung des PüP in die KVNR) zu den Klartextangaben bzw. dem/der Versicherten nur in bestimmten Fällen erfolgen darf, beispielsweise zur Prüfung auf Doppelverdachtsfälle (ähnlich §11 des Landeskrebsregistergesetzes NRW [35]) oder bei geplantem Anschreiben von Versicherten nach Selektion in den medizinischen Daten des FDZ Gesundheit.

Empfehlungen zum Linkage über das FDZ Gesundheit

Nach Überzeugung der Gutachter, wie auch der interviewten Forschenden liegt das weitaus größte Potenzial im Linkage von GKV-Daten mit Forschungsdaten und/oder mit anderen versorgungsnahen Daten. Kausale Schlussfolgerungen sind ausschließlich auf GKV-Routinedaten in aller Regel nicht zu treffen. Stattdessen ergänzen sich GKV-Routinedaten und epidemiologische Daten, Registerdaten und primäre Studiendaten, so dass durch Record-Linkage „das Beste aus beiden Welten“ zusammengebracht werden kann. Umweltbezogene Expositionsdaten, Daten zum Verhalten, Soziodemographie, Lebensqualität und viele weitere je nach Fragestellung wichtige Informationen fehlen in GKV-Routinedaten und müssen durch Record-Linkage zusammengeführt werden, um zentrale wissenschaftliche Fragestellungen angemessen beantworten zu können. Wir sehen daher die Linkagemöglichkeit der GKV-Daten mit anderen Datenquellen als eine entscheidende Voraussetzung für den Erfolg des FDZ Gesundheit.

Wir sehen in der Etablierung praktikabler Verfahren eines Record-Linkage mit GKV-Routinedaten über das FDZ Gesundheit ein sehr großes Potential für die Verbesserung der Qualität, Sicherheit und Wirtschaftlichkeit der Patientenversorgung. Die NAKO zeigt, dass die Akzeptanz der Bevölkerung zur Nutzung und Linkage von GKV-Routinedaten mit

Primärforschungsdaten groß ist. Über 95% der NAKO-Teilnehmenden haben in die Nutzung und das Linkage mit ihren Krankenkassendaten eingewilligt. Für die Umsetzung des Record-Linkage ist eine gut funktionierende Vertrauensstelle entscheidend. Die Vertrauensstelle muss strikt getrennt von den medizinischen Daten operieren. Wir teilen die Einschätzung der Ethikerinnen und Ethiker, dass Identitätsdaten wie Namen/Adressen der KVNR als personenidentifizierende Angabe [65] datenschutzrechtlich gleichgestellt sind. Deshalb führt es nicht zu einer wesentlichen Änderung der Risikobewertung, wenn die Vertrauensstelle am RKI sowohl die KVNR als auch weitere Identitätsdaten hält. Gleichzeitig entstehen aber dadurch bessere Möglichkeiten zur Umsetzung von Linkagestudien. Wir empfehlen daher, analog zum Vorgehen in den Krebsregistern nach §65c SGB V, an der Vertrauensstelle DaTraV die KVNR, Namen, Geburtsdaten und Adressdaten der Versicherten vorzuhalten und prospektiv durch jährlichen Abgleich mit den Einwohnermelderegistern zu pflegen. Die Einwilligung der Studienteilnehmenden zur Speicherung der entsprechenden Daten in der Vertrauensstelle müsste im Informed Consent transparent beschrieben werden.

Durch die Einführung des Broad Consent der MII wird eine bislang einzigartige Möglichkeit für Linkagestudien geschaffen. Beim Broad Consent handelt es sich ebenfalls um einen Informed Consent (informierte Einwilligung), nur dass die Einwilligung zu einer breiteren Nutzung der Daten und nicht nur bezogen auf spezifische Fragestellungen erfolgt. Bisher willigten rund 85% der angefragten Patientinnen und Patienten der Universitätsmedizin im Rahmen des Broad Consent gleichzeitig in das Modul zum Linkage der Klinikdaten mit GKV-Routinedaten ein („Kassen-Modul“). Schon nach kurzer Zeit liegen an den Universitätskliniken in Deutschland aktuell bereits mehr als 70.000 Einwilligungen für das Linkage mit GKV-Daten vor. Das Linkage der GKV-Daten mit den Daten der MII auf Basis des entsprechenden Moduls des Broad Consent sollte über das FDZ umgesetzt werden.

Anders als bei Studien mit einmaligem Linkage müssen bei nationalen Forschungsprojekten wie der MII kontinuierlich (z. B. quartalsweise) linkbare GKV-Routinedaten zur Verfügung gestellt werden. Diese Datenlieferungen sollten in derartigen Fällen vom FDZ Gesundheit an einen zwischengeschalteten Datenhalter erfolgen (im Beispiel Broad Consent der MII die NUM/MII-Infrastruktur), der wiederum fragestellungsbezogen die jeweils notwendigen Daten an die Forschenden herausgibt. In derartigen Fällen ist aufgrund der geringeren Spezifik der Einwilligung eine zusätzliche Governance als Sicherungsmaßnahme notwendig. Diese Verantwortung sollte nicht beim FDZ Gesundheit, sondern bei der Infrastruktur des Forschungsprojekts liegen. Wir schlagen hierzu die Einrichtung von Use & Access Boards vor, die multiperspektivisch besetzt sein sollten und in denen auch klinische Experten und Vertreter der betroffenen Patientinnen und Patienten Sitz und Stimme haben. Für die Datenübergabe aus dem FDZ sollte grundsätzlich ein Vertrag zwischen dem FDZ und der Institution des Antragstellers geschlossen werden.

Grundsätzlich sollte die Übergabe von Daten aus dem FDZ im Rahmen eines Linkage auf Basis einer informierten Einwilligung erfolgen, so wie in dem Beispielfall des Gutachtens (Kohorte NAPKON-POP), der NAKO oder des Broad Consents der Medizininformatik Initiative.

In Ausnahmefällen sollten nach Einschätzung der Gutachter, wie auch der interviewten Krankenkassenvertretenden und Forschenden auch individuelles Datenlinkage mit FDZ-Daten ohne informierte Einwilligung möglich sein. Ein Linkage ohne informierte Einwilligung sollte in bestimmten Fällen ermöglicht werden, beispielsweise

- wenn die Einholung einer Einwilligung nicht bzw. nicht mit vertretbarem Aufwand möglich ist,
- wenn für die Nutzung der Daten bereits eine Einwilligung vorliegt, das Datenlinkage jedoch nicht explizit einbezogen wurde,
- wenn die Studienergebnisse durch einen Selektionsprozess eines Einwilligungsverfahrens maßgeblich an Validität verlieren würden oder
- wenn ein beträchtliches gesellschaftliches Nutzenpotenzial aus der Forschung wahrscheinlich ist.

Die aufgeführten Aspekte sind nicht als vollumfängliche Liste zu verstehen, sondern listen wesentliche Entscheidungsgründe für das Datenlinkage ohne informierte Einwilligung auf.

Ein Beispiel betrifft das Linkage von GKV-Routinedaten mit Daten der Klinischen Krebsregister (KKR). Das vom Innovationsfonds geförderte Projekt WIZEN zeigte, dass ein probabilistisches Linkage von GKV-Daten und den Daten ausgewählter KKR anhand pseudonymisierter KVNR umgesetzt werden kann. Die hierbei notwendig auftretenden Homonym- und Synonymfehler schließen das probabilistische Linkage für wichtige Forschungsfragen jedoch aus. Hierzu gehören Fragen der onkologischen Qualitätssicherung, PatientInnensicherheit und Patientensicherheit, Untersuchung seltener Erkrankungen etc.. Ein momentanes Hindernis für die grundsätzliche Ermöglichung eines Linkages bundesweiter KKR-Daten mit GKV-Daten ist die Uneinheitlichkeit der Regelungen in den unterschiedlichen Landeskrebsregistern. Hier ist dringend eine Harmonisierung notwendig, so dass der große Datenschatz, der durch die Verbindung von GKV-Routinedaten und KKR-Daten entsteht, beispielsweise für die Ableitung von Leitlinienempfehlungen nutzbar wird und damit die medizinische Versorgung von krebserkrankten Menschen in Deutschland wirksam verbessern kann. Weitere Bedarfe für das individuelle Linkage ohne Einwilligung betreffen bspw. Fragen der Pharmakovigilanz und der Pandemic Preparedness. Erlaubnistatbestände für ein Linkage ohne informierte Einwilligung sollten im Gesundheitsdatennutzungsgesetz festgelegt werden, ebenso wie ein transparentes, planbares Prüfverfahren.

Für die Standardisierung und Qualitätssicherung von Forschungsprojekten mit Consent-basiertem Linkage von GKV-Routinedaten und Primärdaten sollten Blaupausen beispielsweise für Studienprotokolle, Informations- und Aufklärungs-, und Einwilligungsdokumente erstellt werden. Durch die Nutzung erprobter Musterdokumente wird der wissenschaftliche Standard der Datennutzung stetig verbessert und Qualitätseinschränkungen immer weiter minimiert.

Diese Blaupausen müssen sich auch an der jeweils geplanten Datennutzung sowie der jeweils aufgebauten Infrastruktur des jeweiligen Forschungsprojekts bzw. -vorhabens orientieren. Dabei ist grundsätzlich zu unterscheiden, ob die Datennutzung konsentbasiert oder ohne Einwilligung erfolgt. Der konsentbasierte Zugang muss zusätzlich unterscheiden, ob eine projektspezifische Einwilligung oder ein Broad Consent vorliegt (siehe hierzu auch Kapitel 2.3). Bei den vorzusehenden Möglichkeiten der wissenschaftlichen Nutzung von Daten unter Beteiligung des FDZ Gesundheit sind weitere Varianten vorzusehen, die sich an der Komplexität der Forschungsfrage sowie der Datenstruktur orientieren sollten. Diese determiniert, welche Variante der Datenbereitstellung innerhalb der Forschungsstruktur zu bevorzugen ist. Im Regelfall sollte bei Vorliegen einer individuellen informierten Einwilligung eine Datenübergabe an das Forschungsinstitut erfolgen. In speziellen, dafür geeigneten Fällen kommt eine Nutzung im FDZ, eine Remote-Nutzung oder verteiltes Rechnen in Betracht.

Datenlinkage

Im Folgenden soll nunmehr davon ausgegangen werden, dass in der Studienstelle die Studierendaten zusammen mit der KVNR vorliegen. Die KVNR ist dabei entweder direkt im Rahmen der Einwilligung erfasst worden oder durch eine der beschriebenen Prozesse nacherhoben worden. Ohne Vorliegen der KVNR ist ein Datenlinkage zwischen Studiendaten und Daten des FDZ Gesundheit nicht möglich.

Die Übermittlung der KVNR an die FDZ-Vertrauensstelle erfolgt in ähnlicher Weise wie in den Prozessen des FDZ für die Krankenkassen beschrieben. Die Studienstelle erzeugt eine Liste der KVNR aller Probanden und versieht diese mit einem Lieferpseudonym. Diese Liste wird an die FDZ-Vertrauensstelle sicher transportverschlüsselt übermittelt. Aus den KVNR wird hier das Periodenübergreifende Pseudonym (PüP) erzeugt und zusammen mit dem Lieferpseudonym an das FDZ Gesundheit übermittelt. Diese Pseudonyme erlauben bei korrekter Handhabung, Daten des FDZ Gesundheit valide mit anderen Datenquellen zu verlinken (konkret siehe Kapitel 4.2). Grundsätzlich sind zwei Szenarien möglich:

Auswertungsszenario A

Das FDZ Gesundheit selektiert die Routinedaten des Versicherten mit Hilfe des temporären Lieferpseudonyms aus dem Gesamtbestand des FDZ. Umfang der Daten, also insbesondere Tabellen, Variablen, zeitlicher Bezug, Struktur der Dateien sind im Vorfeld projektspezifisch

abgestimmt, und im Rahmen des Use & Access-Prozesses geprüft und genehmigt worden (siehe hierzu Kapitel 4.2).

Das FDZ Gesundheit übermittelt die selektierten Daten zusammen mit dem temporären Lieferpseudonym an die anfordernde Stelle mit Hilfe eines sicheren Übertragungsweges (siehe hierzu Kapitel 4.2).

Die anfordernde Stelle übernimmt die Daten und importiert diese in die projektspezifische Studiendatenbank, in der sich auch die Primärdaten des/der Versicherten befinden. Mit Hilfe des temporären Lieferpseudonyms können die Daten nun verlinkt und im Hinblick auf die Forschungsfrage analysiert werden. Nach der wissenschaftlichen Nutzung für die Bearbeitung der vorgesehenen Fragestellungen werden die Daten zunächst beim Forschungsinstitut/der anfordernden Stelle archiviert und gemäß Guter Epidemiologischer Praxis für 10 Jahre für eine mögliche Replikation von Studienergebnissen aufbewahrt. Gemäß Guter Epidemiologischer Praxis müssen Rohdaten bis 10 Jahre nach Projektende verfügbar sein, um gemäß Guter Wissenschaftlicher Praxis und Guter Epidemiologischer Praxis die Überprüfung und Reproduktion von Ergebnissen zu ermöglichen. Eine Nutzung für nicht genehmigte Zwecke darf nicht erfolgen. Nach Ablauf der notwendigen Archivierungsdauer erfolgt eine Löschung der Daten durch die anfordernde Stelle. Dieser Vorgang wird dokumentiert und die erfolgte Löschung sollte dem FDZ von der anfordernden Stelle schriftlich mitgeteilt werden.

In der gesetzlichen Grundlage des FDZ nach § 303 a-f SGB V sowie der Datentransparenzverordnung ist die Funktion eines Datenlinkages über eine Referenzliste sowie die Datenbereitstellung als Export für die Forschung bisher nicht enthalten. Nach Ansicht der Gutachter sowie auch in der Sicht der Ethik (siehe oben) ist hier ausschlaggebend der in der Einwilligung geäußerte Patientinnen und Patientenwille zur Nutzung seiner bzw. ihrer GKV-Routinedaten. Diese Vorgehensweise ist bei Forschungsprojekten in Kooperation mit einzelnen Krankenkassen seit langem üblich. Da die Kooperationsbereitschaft der Krankenkassen, wie oben beschrieben sehr unterschiedlich ist und zudem von der jeweiligen Forschungsfrage abzuhängen scheint, ist die Funktion des FDZ Gesundheit als Datenbereitstellende Institution eine forschungsfreundliche und erwünschte Alternative.

Auswertungsszenario B

Die anfordernde Stelle selektiert aus ihrer Studiendatenbank die Primärdaten der/der jeweiligen Versicherten und übermittelt diese - über einen sicheren Übermittlungskanal - an das FDZ Gesundheit. Das FDZ Gesundheit importiert die Primärdaten in eine projektspezifische Datenbank im FDZ und ergänzt diese in einem zweiten Arbeitsschritt durch die GKV-Routinedaten aus dem Gesamtbestand des FDZ. In einer Studiendatenbank innerhalb der Infrastruktur des FDZ können anhand des temporären Lieferpseudonyms die Daten der

anfordernden Stelle mit denen des FDZ verknüpft und für die Analyse in einer sicheren Umgebung bereitgestellt werden.

Die Analyse der Daten erfolgt in dieser Option durch Forschende per personalisierter sicherer VPN-Verbindung. So genannte Gastarbeitsplätze (GAP) wie sie im DIMDI/DaTraV in Einzelfällen bereitgestellt wurden, sind nach Auskunft des FDZ Gesundheit aktuell nicht geplant.

Vergleich der Szenarien

Beide Szenarien erlauben im Grundsatz eine Analyse der verlinkten Daten. Insbesondere bei der VPN-Verbindung in Szenario B ist die für den Mitarbeiter/für die Mitarbeiterin anfallende Arbeitszeitbelastung identisch, da aufwändige Reisen von der Studienstelle zum FDZ (oder alternativ einer Außenstelle) entfallen.

Aus sicherheitstechnischer Sicht ergeben sich ebenfalls keine großen Unterschiede. Die verlinkten Daten befinden sich jeweils in einer projektspezifischen Studiendatenbank. Zugriff haben ausschließlich die benannten Mitarbeitenden. Datenschutzkonzepte für den Betrieb von projektspezifischen Datenbanken auf Basis von GKV-Routinedaten mit und ohne Verlinkung mit Primärdaten existieren in zahlreichen Projekten, unter anderem in Projekten des Innovationsfonds. Die hierfür erarbeiteten Datenschutzkonzepte sind unter anderem Bestandteil der Datennutzungsanträge (nach SGB X §75) und werden von den zuständigen Aufsichtsbehörden im Zuge des Antragsverfahren geprüft. Diese existierenden Datenschutzkonzepte können als Blaupause für die Implementierung der hier beschriebenen Szenarien dienen.

Aus analytischer Sicht hingegen muss beachtet werden, dass jede Analysestelle spezifische Auswertungsprogramme, Auswertungsumgebungen und Skripte besitzt, mit Hilfe derer die Analyse erfolgt. Eine Portierung dieser Konstellation auf den Studienserver des FDZ kann sich im Einzelfall als technisch anspruchsvoll erweisen und ggfs. mit bestehenden IT-Datenschutz- und Datensicherheitsvorgaben kollidieren². Es ist daher absehbar, dass nicht alle in der Forschungspraxis relevante Konstellationen im FDZ umgesetzt werden können. Dies wird mit großer Wahrscheinlichkeit vor allem solche Konstellationen betreffen, die komplexe und innovative Auswertungen unter Verwendung von bspw. Bilddaten, Labordaten, -omics Daten, sowie von Daten mit spezifischen Formaten oder Eigenschaften betreffen, die in den datenführenden Einrichtungen durch spezifische Datenbank-Implementierungen und -

² Aus Erfahrung des Gutachterkreises gestaltete sich die Implementierung einer spezifischen Auswertungsumgebung für GKV-Routinedaten in zwei Krankenkassen bzw. deren IT-Dienstleistern als sehr zeit- und arbeitsaufwändig. Diese Einschränkungen bestätigten sich auch im Umgang mit weiteren Forschungsdatenzentren öffentlicher Datenproduzenten. Als Vorteil kann die Nutzung bereits validierter Auswertungsmodule/-skripte und die erweiterte Funktionalität spezifischer Umgebungen benannt werden.

konfigurationen abgebildet sind (z. B. relationale Strukturen, hierarchische Strukturen), die nur mit großem Aufwand portierbar sind.

Auf der Seite des FDZ wären für die Aufnahme der jeweiligen Kohortendaten die notwendigen Infrastrukturen vorzuhalten. Diese sind zwischen verschiedenen Registern und Kohorten heterogen und eine Umwandlung in ein bestimmtes oder eines von wenigen Datenbankformaten wird in vielen Anwendungsfällen nicht einfach möglich sein. Im Regelfall stellt der Datenbestand der anfordernden Stelle die größere Herausforderung dar - exportierte FDZ-Daten können erheblich besser in eine andere Datenbankumgebung überführt werden, als eine Vielzahl von heterogenen Dateninfrastrukturen in eine oder wenige vom FDZ Gesundheit angebotenen Umgebungen. Erhebliche Zeitverluste im Forschungsprozess sowie hohe individuelle Beratungsbedarfe auf der Einzelprojektebene werden eine häufige Folge sein.

Unter dem Primat der Datensparsamkeit stellt sich die Frage der Verhältnismäßigkeit, da in den meisten Fällen der Variablenbestand in den anfordernden Stellen deutlich höher ist, als der vom FDZ Gesundheit beantragte Variablensatz. Hier kommt es u. U. zu Kapazitätsproblemen in der Infrastruktur des FDZ., wenn bspw. umfangreiche Bilddaten oder -omics Daten übertragen werden müssen.

Aus Sicht der Datensicherheit muss kritisch diskutiert werden, dass durch Überführung von Fremddatenbeständen an das FDZ Gesundheit dort größere Sammlungen an Primärdaten entstehen würden, die ein Angriffsziel bilden können. In den Infrastrukturen der Einzelprojekte erfolgt die Archivierung von Rohdaten aus dem FDZ nur so lange, wie dies die Gute Epidemiologische Praxis gebietet. Danach erfolgt eine unverzügliche, zu dokumentierende Löschung der Daten in der jeweiligen Infrastruktur des Einzelprojektes. Das FDZ sollte über die erfolgte Löschung schriftlich informiert werden. Die Überführung ihrer Daten an das FDZ ist möglicherweise nicht im Interesse aller Studienteilnehmenden. Ein Anreiz für Personen zur Studienteilnahme ist häufig auch die Unterstützung von Forschung und Forschungsinstitutionen mit besonderem Schwerpunkt, für den sich die Teilnehmenden engagieren wollen (z. B. Forschung zu Krebs, zu Neurodegenerativen Erkrankungen oder zu Allergien). Die Studienteilnehmenden dokumentieren mit ihrer informierten Einwilligung auch das Vertrauen in die Datenmanagementprozesse der durchführenden Einrichtung. Die Überführung von Forschungsdaten am FDZ Gesundheit und die dortige Speicherung erfordert in allen laufenden Studien und Registern eine Erweiterung der informierten Einwilligung. Dies könnte mit einer geringeren Teilnahmequote assoziiert sein, weil das FDZ Gesundheit als neutrale staatliche Institution eben nicht für das Engagement in einem bestimmten Bereich steht, so wie andere Institutionen oder Studiengruppen. Dem könnten ggf. Löschkonzepte entgegenwirken

Um die Anforderung der Guten Epidemiologischen Praxis zu erfüllen, müssen alle Ergebnisse mindestens für 10 Jahre nach der Publikation nachrechenbar sein. Daraus folgt, dass bei einer Übertragung der Primärdaten an das FDZ dieses die Auswertedatensätze für mindestens 10 Jahre speichern muss. Eine kurzfristige Löschung unmittelbar nach erfolgter Auswertung würde dazu führen, dass die GEP Anforderungen nicht erfüllt werden.

Datenschutzrechtlich müssten im typischen Fall eine Vielzahl von patientenbezogenen hochinformativen Daten aus den anfordernden Einrichtungen an das FDZ Gesundheit übertragen werden. Die Datenschutzverantwortlichkeit der übertragenen personenbezogenen Daten ginge in jedem Fall auf das FDZ über. In der Kombination mit den zugeordneten GKV-Daten steigt das Re-Identifikationsrisiko im FDZ Gesundheit insbesondere bei dauerhafter, aber grundsätzlich auch bei temporärer Speicherung der zugeordneten GKV-Daten. Dieses zusätzliche Risiko müsste in der Datenschutzfolgeabschätzung und auf allen administrativen und technischen Ebenen des FDZ berücksichtigt und ggfs. durch entsprechende zusätzliche Maßnahmen minimiert werden.

In allen Fällen, in denen kommerzielle Software für die Analyse eingesetzt wird, sind lizenzrechtliche Fragen zu klären. So kann nicht vorausgesetzt werden, dass eine (preiswerte) universitäre Lizenz für ein Statistikpaket auf einen außerhalb der Universität stehenden Server übertragen werden darf. Die Sicherheitsabteilung des FDZ ihrerseits muss umfänglich klären, welche Voraussetzungen geschaffen werden müssen, um beispielsweise Eigenentwicklungen/Skripte des externen Mitarbeiters auf der FDZ-Infrastruktur installieren zu können. Dies gilt perspektivisch auch für den Standardfall FDZ-Datennutzung.

Die Gutachter sprechen sich zusammenfassend deutlich für Szenario A als Regelfall aus. Szenario B könnte alternativ in geeigneten Fällen angeboten werden, darf jedoch nicht die vordringliche und keinesfalls die einzige Umsetzungsmöglichkeit darstellen.

Alternative Auswertungsszenarien

Neben den beiden genannten Szenarien werden häufig unter dem Stichworten "verteiltes Rechnen", "code to the data", "Auswertungswizards" (siehe beispielsweise [66]) weitere technische Lösungen diskutiert. Leitgedanke ist hierbei, dass die physische Zusammenführung von Datenbeständen nicht in allen Fällen erforderlich ist. Unter datenschutzrechtlichen Aspekten stellt dies sicherlich einen interessanten Ansatz dar. Die Nutzbarkeit solcher Verfahren hängt allerdings stark von dem konkreten Anwendungsfall ab. Einfache Mittelwertberechnungen benötigen in vielen Fällen keine Zusammenführung von Datenbeständen und sind damit anders zu bewerten als beispielsweise Multilevelanalysen, die je nach Detailgrad der benötigten Daten und Verfahren einen zusammengeführten Gesamtdatensatz erfordern. Die Durchführung von Multilevel-Analysen unter Anwendung deterministischer Methoden ist in verteilten Datensätzen aktuell nicht möglich [54]. Ein

vergleichsweise neues Verfahren, föderiertes Lernen basierend auf maschinellem Lernen (künstlicher Intelligenz), gewinnt im Bereich der Gesundheitsforschung aktuell an Bedeutung. Mit diesem Verfahren könnte die Durchführung von Multilevel-Analysen zukünftig möglich werden. Die wenigen pilothaften Anwendungen zeigen aber bislang im Vergleich zur Analyse zentraler Datensätze ungenaue oder fehlerhafte Parameterschätzungen [67-70]. Gerade die Präzision der relevanten statistischen Parameter ist im Bereich der Gesundheitsforschung von höchster Priorität. Fehlerhafte Schätzungen können fatale Konsequenzen nach sich ziehen [71]. Daher ist eine Überführung von Methoden des föderierten Lernens in die Praxis zur Anwendung im Bereich der Gesundheitsforschung aktuell nicht sicher. Zum gegenwärtigen Zeitpunkt ist daher keine Alternative zur Zusammenführung zu einem zentralisierten Datensatz verfügbar und damit eine zentrale Datenspeicherung und -auswertung für die Untersuchung relevanter Fragen der Versorgungsforschung unabdingbar. Ob dies in einigen Jahren anders werden wird, ist gegenwärtig nicht vorauszusagen.

Aus Sicht der Gutachter ist ein dezentraler Ansatz, bei dem die Kassendatenbestände nicht zu einem zentralen Datensatz zusammengeführt werden würden, zum jetzigen Zeitpunkt keine Alternative zu den oben beschriebenen Szenarien. Dies gilt insbesondere vor dem Hintergrund des Detailreichtums von klinischen Primärdaten und GKV-Routinedaten. Die weitere Entwicklung dieses Forschungsfeldes sollte jedoch für zukünftige Implementierungen weiterhin beobachtet werden, und die Nutzbarkeit sollte jeweils begleitend an konkreten Anwendungsfällen validiert werden.

Empfehlung

Im Regelfall ist die anfordernde Stelle eine Kohorte oder ein Register. Hier werden hochinformativ personenzugehörige Daten gehalten, häufig mit umfangreichen Variablenansätzen. Die Datenbankstrukturen sind heterogen, ebenso die verwendeten Auswerteprogramme und -routinen. Eine Anpassung an vom FDZ bereitgestellte Datenbank-Infrastruktur ist aufwändig und nicht in allen Fällen möglich. Ein Vorteil hinsichtlich des Datenschutzes resultiert nicht, da die Daten des FDZ-Gesundheit beim Export in die Dateninfrastruktur der anfordernden Kohorten stets in eine Umgebung überführt werden, die für die Haltung und prospektive Erweiterung hochinformativer personenzugehöriger Daten geprüft und sowohl ethisch als auch datenschutzrechtlich unbedenklich ist. Die rechtliche Grundlage der Datenverarbeitung sollte jeweils ein Vertrag zwischen der durchführenden Institution und dem FDZ sein.

Der administrative Aufwand für das FDZ ist in der Export-Variante minimal. Die Datenschutzverantwortlichkeit liegt bei der anfordernden Stelle. Umgang, Löschvorschriften, Nicht-Weitergabe etc. kann in Verträgen verbindlich geregelt werden, für die Vorlagen und Erfahrungen in zahlreichen Zentren in Deutschland, teilweise schon seit mehr als 15 Jahren und über mehrere tausend erfolgreich durchgeführte Datenübergaben existieren. Wir

empfehlen, dass Szenario A der Regelfall wird. Szenario B kann in Fällen erwogen werden, in denen die Datensätze der anfordernden Stellen wenige Variablen in einer üblichen, z. B. SQL-Umgebung ohne komplexe Relationalstruktur umfassen.

Die temporäre Übernahme externer Datenbestände in das FDZ wirft erhebliche Datensicherheits-, Datenschutz- und lizenzrechtliche Fragen auf. Im konkreten Fall ist mit einem erheblichen Beratungsbedarf durch das FDZ zu rechnen, der zusätzliche Personalressourcen erfordert und über die Dauer des jeweiligen Auswertungsprojektes bindet.

4.1.2 Empfehlungen zum Identitätsmanagement

Für die unter Kapitel 2 genannten Anwendungen, welche künftig auch, wenn nicht vor allem mit Einbeziehung des FDZ Gesundheit realisiert werden sollen, ist eine validierte und plausibilisierte Datenbasis notwendige Voraussetzung. Dabei sind einzelne Aspekte wie Harmonisierung der Daten oder Vollständigkeitsprüfungen eher unter technischen Gesichtspunkten einzuordnen und mit Hilfe standardisierter Prozesse zu lösen. Herausfordernder ist die Datenvalidierung von Routinedatenquellen, die ein kontinuierlicher Prozess ist und auch Ergebnisse und Erkenntnisse aus inhaltsbezogenen Forschungsprojekten und methodisch orientierten Projekten mit einbezieht. Validierung bedeutet auch, Fehler in der Datenbasis zu beheben bzw. zu minimieren. Allgemein gilt, dass eine geringe Fehlerrate notwendige Bedingung zur Erreichung des jeweiligen Forschungszieles ist. Dies schließt auch ein so genanntes Identitätsmanagement mit ein, also die Frage, ob die mit dem periodenübergreifenden Pseudonym gespeicherten Daten des FDZ Gesundheit tatsächlich genau einer und derselben Person entsprechen.

Der aktuelle Datenfluss des FDZ Gesundheit sieht nach Kenntnis der Gutachter keine Stelle vor, an der ein solcher Identitätsabgleich durchgeführt wird. Identifikator für einen Versicherten bzw. eine Versicherte ist die KVNR bzw. ein hieraus gebildetes Pseudonym. Der Datenfluss von den Krankenkassen in das FDZ Gesundheit legt nahe, dass diese Prozesse nicht fehleranfällig für die korrekte Zuordnung der Daten bzw. für die Datenlinkage selbst ist. Die Frage nach der Eindeutigkeit dieses Identifikators ist daher bei der Vergabe der KVNR selbst unter Bewertung der vorher ablaufenden Prozesse zu beantworten. Da die KVNR nach einem eindeutigen - technisch nicht fehlerbehafteten - Verfahren aus der Rentenversicherungsnummer (RVNR) generiert wird, muss die Prüfung bei der Eindeutigkeit des Identifikators RVNR ansetzen. Dass es bei der Vergabe der RVNR selbst zu Fehlern kommen kann, die dann auch Auswirkung auf die vergebenen KVNR hat, wird in einem Dokument des GKV-Spitzenverbandes thematisiert [72]. Konkret wird ein fehlerhaftes Geburtsdatum, die Vergabe von mehr als einer Nummer für eine Person oder die Vergabe einer RVNR an mehrere Personen als mögliche Fehler benannt. In solchen Fällen werden, sofern diese bekannt werden, fehlerhafte Rentenversicherungsnummern gesperrt und neue korrekte RVNR

vergeben. Um die Auswirkung einer solchen fehlerhaften Vergabe auf das KVNR-basierte System der GKV-Routinedaten bewerten zu können, bedarf es einer Kenntnis über die Anzahl der bisher entstandenen Fehler bzw. der neu vergebenen RVNR. Unter Kenntnis einer solchen Fehler- bzw. Neuvergabehistorie könnten im FDZ Gesundheit die bisher verfügbaren Routinedaten geprüft, Korrekturen durchgeführt oder Daten mit fehlerhaft behafteten RVNR bzw. KVNR gesperrt werden.

Ebenso kann und muss an dieser Stelle die Frage nach der Eindeutigkeit des Stammverzeichnisses der Rentenversicherung gestellt werden. Welche qualitätssichernden Schritte hierfür notwendig sind und welche davon routinemäßig durchgeführt werden, konnte im Rahmen dieses Gutachtens nicht eruiert werden. Es wird empfohlen, diese Prozesse transparent darzustellen und im Hinblick auf die Eindeutigkeit der RVNR bzw. der daraus abgeleiteten KVNR auf den Prüfstand zu stellen.

Von einer Vertrauensstelle, die für das Identitätsmanagement verantwortlich zeigt, sind regelhaft die im Folgenden aufgeführten Aufgaben zu leisten:

- Prüfung der Eindeutigkeit von Identifikatoren, also Vermeiden bzw. Minimierung von Homonym und Synonymfehlern
- Verlinkungsmöglichkeit über personenidentifizierende Klartextangaben
- Anschreiben von Personen mit der Möglichkeit, Daten bei Betroffenen zu prüfen, zu korrigieren und/oder nachzuerheben.
- Dokumentation der Vergabe von KVNR und RVNR, insbesondere bei Neuvergabe dieser Identifikatoren in Fällen fehlerhafter Vergabe (siehe oben)

Erläuterung: Die Prüfung der Eindeutigkeit ist eine qualitätssichernde Maßnahme, die im Rahmen des Aufbaus und Validierung einer Datenquelle kontinuierlich durchzuführen ist. Das Vorhalten und die Pflege der personenidentifizierenden Klartextangaben der Versicherten ermöglicht die Verlinkung von Datenkörpern, die den notwendigen Identifikator Krankenversicherungsnummer (KVNR) nicht erfasst haben und bei den Probanden/bei der Probandin auch nicht nacherfassen können.

Für die Möglichkeit, Personen anzuschreiben und nachträglich Daten zu erheben soll folgendes Beispiel skizziert werden: Im FDZ Gesundheit können Personen mit auf Basis der dokumentierten Diagnoseangaben selektiert werden. Die Datenbasis mit mehr als 70 Millionen Versicherten erlaubt es in einzigartiger Weise, auch für seltene Erkrankungen (weniger als 5 von 10.000 Menschen betroffen) hinreichend große Fallzahlen zu generieren und für diese Personengruppe relevante in den GKV-Abrechnungsdaten nicht vorhandene Parameter projektspezifisch nachzuerheben. Für das Anschreiben von Personen ist sicherzustellen, dass im Versichertenverzeichnis der Vertrauensstelle jeweils die aktuellen Wohnort-

daten enthalten sind. Dafür ist ein kontinuierlicher Abgleich (z. B. monatlich) mit den Versichertenverzeichnissen der Krankenkassen bzw. der Datenstelle Rentenversicherung erforderlich. Fehlerbehaftete RVNR und daraus ebenfalls fehlerbehaftete KVNR können zu Problemen bei der Analyse langer Beobachtungszeiträume führen, wenn Versicherte lediglich aufgrund der Neuvergabe einer KVNR nicht mehr durchgehend longitudinal beobachtet werden können. Eine vollständige Dokumentation und ggfs. auch ein Mapping von alten auf neue KVNR könnte dieses Problem beheben.

Es wird empfohlen, in der Vertrauensstelle DaTraV eine Datenbank mit personenidentifizierenden Klartextangaben der Versicherten und Bezug zur Krankenversicherungsnummer anzulegen. Die Erfassung dieser Daten könnte entweder durch Lieferung der in den Krankenkassen dokumentierten Personenangaben (§ 288 SGB V, Versichertenverzeichnis) oder durch Import und kontinuierlichen Abgleich der Daten mit denen der Datenstelle der Rentenversicherung (§145ff SGB VI). Entsprechende gesetzliche Grundlagen sind zu schaffen. Alternativ können die beschriebenen Funktionen auch an die Datenstelle Rentenversicherung übertragen werden. Hierdurch könnten Parallelstrukturen, also das Vorhalten der Versichertenverzeichnisse in zwei Vertrauensstellen mit der Notwendigkeit, den Datenstand wechselseitig abgleichen zu müssen, vermieden werden.

4.1.3 Rechtliche/Juristische Empfehlungen zum FDZ Gesundheit

Nach dem Wortlaut des § 75 SGB X ist die Einwilligung eine von mehreren Voraussetzungen für die Zulässigkeit der Übermittlung von Sozialdaten: Weitere Voraussetzungen sind ein zulässiger Zweck, Interessenabwägung und ein adäquates Datenschutzkonzept. Bei vorliegender informierter Einwilligung und positivem Ethikvotum sollte zukünftig ein Antrag nach §75 SGB X für den Datenexport aus dem FDZ Gesundheit an die Studiengruppe nicht notwendig sein, da die Prüfung des zulässigen Zwecks, der Interessenabwägung und des Datenschutzkonzeptes ebenfalls bereits durch die zuständige Ethikkommission erfolgt. Die derzeitige Mehrfachprüfung von Datenschutzaspekten durch nicht miteinander in Verbindung stehende Institutionen wie in diesem Fall Ethikkommissionen und Aufsichtsbehörden von Krankenkassen bringt keinen Vorteil bzgl. Datenschutz und Datensicherheit und führt zu nicht planbaren, unangemessen aufwändigen Verfahren für die Forschenden. Der Wortlaut des §75 SGB X sollte daher im Rahmen des Gesundheitsdatennutzungsgesetzes überarbeitet werden, so dass die Notwendigkeit eines §75-Antrags bei Consent-basierter Linkage von GKV-Daten mit anderen Daten entfällt. Sollte wie im Fallbeispiel NAPKON das Datenlinkage fragestellungsbezogen erfolgen, so sollte eine separate Datenannahme- und -aufbereitungsstelle seitens der Infrastruktur der Studiengruppe vorhanden sein, an die das FDZ Gesundheit die Daten bereitstellt. Im Einklang mit den Empfehlungen der Expertinnen und Experten für Ethik empfehlen wir in diesem Szenario die Einrichtung eines multiperspektivischen Use-and-Access-Committee (UAC) unter Einbeziehung klinischer Expertise und Vertretern der betroffenen Patientinnen und Patienten in der Infrastruktur der Studiengruppe. Dieses UAC

entscheidet über die Datenbereitstellung für die einzelnen Projekte und Fragestellungen der Studiengruppe. Diese Vorgehensweise ist bereits gängige Praxis beispielsweise in den SHIP-Kohorten, der NAKO Gesundheitsstudie, des Deutschen Zentrums für Herz-Kreislaufforschung (DZHK) und des Zentrums für Krebsregisterdaten (ZfKD) des RKI. Kriterien für die Infrastrukturkomponenten, die bei größeren Kohortenstudien mit GKV-Datenlinkage erforderlich sind, sollten definiert und Blaupausen für deren Umsetzung entwickelt und für die Forschungscommunity bereitgestellt werden.

Einwilligungserklärungen zur Nutzung und zum Linkage von GKV-Routinedaten sind je nach Projekt z. T. spezifisch z. B. für eine definierte Fragestellung zur Arzneimittelsicherheit), z. T. breiter (wie in dem Fallbeispiel NAPKON, bei dem ein breites Spektrum an Fragestellungen zur COVID-19 vorgesehen ist) und können im Falle des Broad Consent der MII die medizinische Forschung allgemein betreffen. Generell gilt, dass die Governance Strukturen (Treuhandstelle, Nutzungsordnungen, SOPs, Einbindung der zuständigen Ethik-Kommissionen, UAC, von Auswertestelle getrennte GKV-Datenannahme- und Aufbereitungsstelle, technische Vorkehrungen bei der Übermittlung von Daten etc.) seitens der Forschergruppe, die die Daten vom FDZ Gesundheit empfängt, umso robuster ausgebildet sein müssen, je breiter die Einwilligungserklärung zur GKV-Datennutzung ist. Dies betrifft die Breite der Daten (Anzahl Variablen) und die Länge des Zeitraums, für den die Studienteilnehmenden in eine Verknüpfung mit FDZ-Daten einwilligen. Patientenvertretende halten je nach Fragestellung die Einwilligung in die Nutzung von zukünftigen GKV-Routinedaten in Abhängigkeit der Forschungsfragestellungen für Zeiträume bis zu 10 Jahren und länger für angemessen. Diese Sichtweise deckt sich mit den Empfehlungen der Gutachter. Essenziell ist in den anfordernden Studien stets die Möglichkeit des Widerrufs der Einwilligung in die Datenverknüpfung zu jedem Zeitpunkt durch die Studienteilnehmenden vorzusehen.

Wie oben ausgeführt, ist ein Datenlinkage von klinischen Primärdaten und Daten des FDZ Gesundheit technisch problemlos möglich, wenn das aus der Krankenversicherungsnummer (KVNR) generierte Periodenübergreifende Pseudonym (PüP) vorliegt. Wo die KVNR in den anfordernden Studien nicht vorhanden ist, muss es möglich sein, dass die KVNR und das PüP über die identifizierenden Klartextangaben des Probanden/der Probandin angefragt werden kann. Grundsätzlich gilt hierbei, dass der Prozess des Datenlinkage, also die Zusammenführung von personenbezogenen Gesundheitsdaten, immer auf Basis von Pseudonymen und damit nicht direkt personenidentifizierbaren Identifikatoren erfolgt. Damit ist sichergestellt, dass die personenbeziehbaren Angaben getrennt von den Gesundheitsdaten, in diesem Fall auch den verlinkten Gesundheitsdaten verarbeitet werden. Keine Person, die im Forschungsprozess mit den verlinkten Gesundheitsdaten arbeitet, hat damit Zugriff auf die direkt personenidentifizierbaren Angaben (Trennungsgebot, siehe hierzu auch Kap. 1.3).

In Übereinstimmung mit den Ansicht der Ethik ist nach Meinung der Gutachter die Patienteneinwilligung als informierte Willensäußerung des Patienten für die Verlinkung der medizinischen Daten des Probanden/der Probandin und deren Übermittlung an den/die Forscher:in ausreichend (Informationelle Selbstbestimmung). In der Rechtsgrundlage des FDZ Gesundheit (§303 a bis f SGB V) sowie der zugehörigen Datentransparenzverordnung findet sich allerdings kein Hinweis auf ein konsentbasiertes Datenlinkage, unabhängig davon, ob das Linkage im FDZ selbst stattfindet (mit Übermittlung der klinischen Primärdaten an das FDZ) oder beim Forscher (nach Export der Daten durch das FDZ Gesundheit). Die Gutachter empfehlen eine entsprechende Anpassung der Datentransparenzverordnung. Wenn die Datenbereitstellung nicht durch das FDZ Gesundheit sondern durch Krankenkassen erfolgt, empfehlen die Gutachter zudem, dass bei Vorliegen einer informierten Willensäußerung des Patienten/der Patientin ein Antrag nach §75 SGB X nicht notwendig ist oder ggfs. das Antragsverfahren durch ein fast-lane Verfahren beschleunigt werden sollte (zur Begründung siehe Kap. 2.7).

4.2 Anwendung auf NAPKON

Das Linkage von Daten verschiedener Krankenkassen mit NAPKON-Studiendaten als Proof-of-Concept wurde bereits unter Kapitel 1.3 detailliert beschrieben. Ein Datenlinkage mit NAPKON-Studiendaten unter Einbeziehung des FDZ Gesundheit als kontinuierlicher Routineprozess würde auf analogen Strukturen basieren, und muss lediglich in einzelnen Punkten konkretisiert werden. Die bestehende Struktur in NAPKON sieht dabei vor, dass die Studiendaten der NAPKON-Kohorten in einer zentralen Stelle (NUKLEUS Dateninfrastruktur) vorgehalten werden. Der im Folgenden beschriebene Prozess geht daher davon aus, dass die Routedaten der NAPKON-Probanden vom FDZ Gesundheit bereitgestellt werden und in die Dateninfrastruktur von NAPKON integriert werden.

Zunächst nimmt das FDZ Gesundheit samt der am RKI angesiedelten Vertrauensstelle die Rolle ein, welche zuvor einzelne Krankenkassen übernommen haben (siehe Abbildung 2). Weiterhin wird die Verknüpfung über die KVNR vorgenommen, welche durch die NUKLEUS Treuhandstelle zur Verfügung gestellt wird. Diese wird für alle Studienteilnehmenden mit Einwilligung in die Datenverknüpfung an die Vertrauensstelle am RKI zusammen mit einem Transport-Pseudonym (TranspPID1) übermittelt. Ein Aufteilen der Angaben aus der Schlüsselvariable KVNR auf verschiedene Krankkassen ist dabei nicht mehr erforderlich, da im FDZ Gesundheit die Daten aller gesetzlichen Krankenkassen vorliegen. In der Vertrauensstelle am RKI wird die KVNR in das Periodenübergreifende Pseudonym (PüP) überführt, sodass anschließend eine Zuordnungstabelle der beiden Pseudonyme PüP und TranspPID1 an das FDZ Gesundheit übermittelt werden kann. Im FDZ Gesundheit wird mit dieser Information aus der vorliegenden Datenbasis (MDAT2) ein auf die genannten NAPKON-Studienteilnehmenden

eingeschränkter Datenbestand extrahiert und das Pseudonym PÜP für diesen Datenbestand durch das Pseudonym TranspPID1 ersetzt (MDAT2_{TranspPID1}). Dieser Teildatenbestand wird anschließend an die Datenannahme- und Aufbereitungsstelle übermittelt. Diese liest die vom FDZ Gesundheit erhaltenen Daten ein und bereitet diese weiter auf - inklusive einer Plausibilitäts- und Konsistenzprüfung. Um eine permanente Datenhaltung der aufbereiteten Daten des FDZ in der Datenannahme- und Aufbereitungsstelle zu vermeiden, wird nach Abschluss der Prüfung der aufbereitete Datensatz verschlüsselt, an die NUKLEUS-Datenhaltung übermittelt und der Datenbestand in der Datenannahme- und Aufbereitungsstelle gelöscht. Der Schlüssel zur Entschlüsselung der Daten verbleibt dabei in der Datenannahme- und Aufbereitungsstelle. So ist sichergestellt, dass anlassbezogen eine Entschlüsselung möglich ist, der Datenbestand aber nicht ohne Interaktion zwischen Datenannahme- und Aufbereitungsstelle und NUKLEUS Datenhaltung genutzt werden kann. Diese Interaktion wird nur initiiert, wenn für ein Forschungsprojekt ein projektspezifischer Datenauszug erstellt werden soll. Ein solcher Fall tritt beispielsweise ein, wenn eine Publikation nach den Anforderungen der Guten Wissenschaftlichen Praxis extern überprüft werden soll. In diesem Fall müssen alle darin enthaltenen Auswertungen aus den Original-Rohdaten von unabhängiger Seite reproduziert werden können. Eine solche Prüfung muss bis zu 10 Jahre nach der Publikation möglich sein.

Der gesamte Selektions- und Exportprozess wird regelmäßig und unabhängig von konkreten Forschungsvorhaben durchgeführt. Mit diesem Verfahren kann auf eine wiederholte Datenlieferung aus dem FDZ und Neuaufbereitung der erhaltenen Daten für jedes einzelne Forschungsprojekt verzichtet werden. Diese Datenlieferung muss immer dann erfolgen, wenn entweder neue Daten im FDZ Gesundheit vorliegen oder die Kohorte der NAPKON-Studienteilnehmenden mit Einwilligung in die Datenverknüpfung sich nachhaltig geändert hat.

Neben dem regelmäßigen Routine-Datenfluss bezüglich der GKV-Daten bis zur Datenannahme- und Aufbereitungsstelle ist noch ein projektspezifischer bzw. anlassbezogener Datenfluss zu etablieren, welcher unabhängig vom FDZ Gesundheit ist. Hier wird analog zum unter Kapitel 1.3 beschriebenen Prozess der Datenfluss und die Erstellung eines projektspezifischen NAPKON-Datenauszugs innerhalb der NAPKON-/NUKLEUS-Struktur (Beteiligung der NAPKON-Kohorten, des NAPKON UAC, der NUKLEUS Transferstelle und der NUKLEUS Treuhandstelle) durchgeführt. Dieser Prozess wird ausgelöst von einer Anfrage für ein Forschungsprojekt basierend auf verlinkten GKV- und NAPKON-Daten. Im Ergebnis steht ein projektspezifischer Datensatz aus Daten der NAPKON-Kohorten (ProjDAT1), wobei die spätere Zuordnung zu den GKV-Daten über ein Pseudonym erfolgt, welches für jedes Forschungsprojekt separat erzeugt wird (TranspPIDx). Die separate Erzeugung stellt technisch sicher, dass keine Angaben über Forschungsprojekte hinweg verknüpft werden können.

Ebenso erfolgt die Weiterverarbeitung der aufbereiteten und in der NUKLEUS Datenhaltung verschlüsselt gespeicherten Daten ausschließlich anlassbezogen jeweils für ein Forschungsprojekt. Für ein durch das UAC genehmigtes Forschungsprojekt werden die verschlüsselten Daten aus der NUKLEUS Datenhaltung an die Datenannahme- und Aufbereitungsstelle übermittelt und dort entschlüsselt. Anschließend wird ein projektspezifischer Datensatz aus Daten (ProjDAT2) extrahiert und das Pseudonym TranspPID1 durch das bereits erwähnte projektspezifische Pseudonym TranspPIDx ersetzt. Dieses projektspezifische Pseudonym, welches in der NUKLEUS Treuhandstelle erzeugt wird, ist identisch zu dem für NAPKON genutzten projektspezifischen Pseudonym und ermöglicht so die anschließende Verknüpfung der projektspezifischen GKV-Daten mit den projektspezifischen NAPKON-Daten. Diese Verknüpfung erfolgt in der Analysestelle, welche beide projektspezifischen Datensätze aus unterschiedlichen Quellen erhält (ProjDAT1 aus NUKLEUS Transferstelle und ProjDAT2 aus Datenannahme- und Aufbereitungsstelle). Nach Übermittlung der projektspezifischen Daten (ProjDAT2) werden die $MDAT2_{TranspPID1}$ in der Datenannahme- und Aufbereitungsstelle wieder gelöscht.

Die Gutachter empfehlen, die Nutzung der verknüpften Datenbestände durch Antragsteller an ein Zertifikat zu binden, welches eine Grundkompetenz der Routinedatennutzung nachweist. Dieses wird unter Kapitel 4.3 näher beschrieben. Nach Verknüpfung der Daten ProjDAT1 und ProjDAT2 anhand des projektspezifischen Pseudonyms TranspPIDx in der Analysestelle kann dort auch die geplante Auswertung der verlinkten Daten vorgenommen werden.

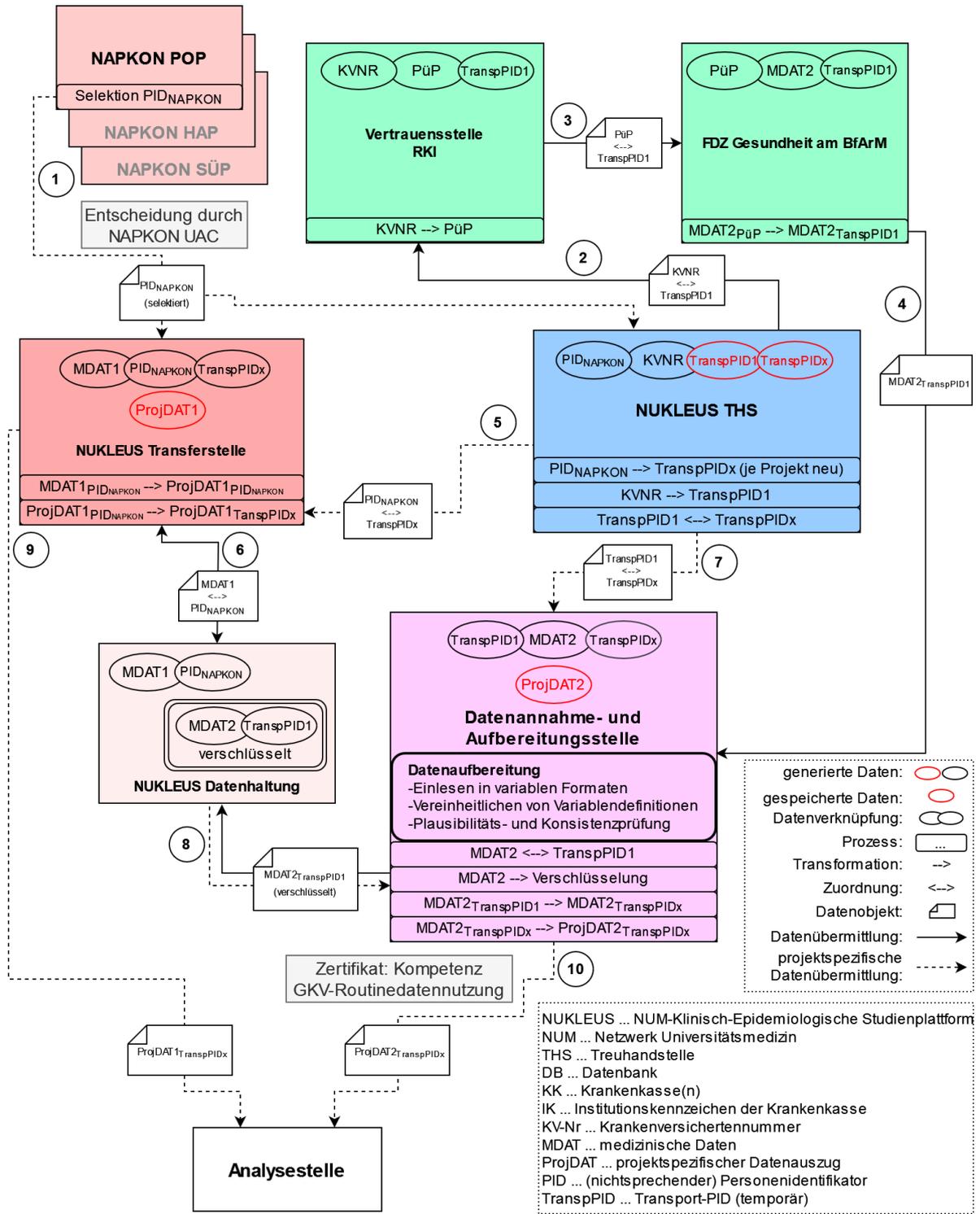


Abbildung 2: Datenfluss, Datenverknüpfung und Datenanalyse - Routineprozess unter Beteiligung des FDZ Gesundheit

Tabelle 6: Legende zu den nummerierten Datenflüssen aus Abbildung 2

Nummer	verantwortliche Stelle	Datenübermittlung
(1)	NAPKON POP	Übermittlung des NAPKON-Pseudonyms von NAPKON-Studienteilnehmenden mit Einwilligung zur Verknüpfung der Studiendaten mit Krankenkassendaten, um diese anschließend in der NUKLEUS Transferstelle und NUKLEUS Treuhandstelle für nachfolgende Datenflüsse selektieren zu können.
(2)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt aus der KVNR ein Transport-Pseudonym (TranspPID1) und übermittelt dieses zusammen mit einer Zuordnung der vorliegenden KVNR für alle NAPKON-Studienteilnehmenden mit entsprechender Einwilligung an die Vertrauensstelle am RKI.
(3)	Vertrauensstelle am RKI	Die Vertrauensstelle am RKI ersetzt die KVNR durch das periodenübergreifende Pseudonym (PüP) und übermittelt die so entstandene Zuordnungstabelle PüP-TranspPID1 an das FDZ Gesundheit.
(4)	FDZ Gesundheit	Das FDZ Gesundheit selektiert die vorliegenden medizinischen Daten der GKV (MDAT2) auf die übermittelten PüP sowie den vorab vereinbarten Datenkranz, ersetzt für diesen Datenauszug die PüP durch das Transport-Pseudonym (TranspPID1) und stellt den so bearbeiteten Datenbestand (MDAT2 _{TranspPID1}) der Datenannahme- und Aufbereitungsstelle am ZEGV zur Verfügung.
(5)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt aus dem NAPKON-Pseudonym von NAPKON-Studienteilnehmenden mit Einwilligung ein projektspezifisches Pseudonym (TranspPIDx) und stellt dieses zusammen mit einer Zuordnung des NAPKON-Pseudonym der NUKLEUS Transferstelle zur Verfügung.
(6)	NUKLEUS Transferstelle und Datenhaltung	Die medizinischen NAPKON-Daten (MDAT1) werden zusammen mit dem NAPKON-Pseudonym in der NUKLEUS Datenhaltung gespeichert, anlassbezogen in der NUKLEUS Transferstelle verarbeitet und zwischen beiden Stellen übermittelt.
(7)	NUKLEUS Treuhandstelle	Die NUKLEUS Treuhandstelle erzeugt für NAPKON-Studienteilnehmende mit Einwilligung eine Zuordnungstabelle zwischen dem projektspezifischen Pseudonym (TranspPIDx) sowie dem Transport-Pseudonym (TranspPID1) und stellt diese der Datenannahme- und Aufbereitungsstelle am ZEGV zur Verfügung.

(8)	Datenannahme- und Aufbereitungsstelle und NUKLEUS Datenhaltung	Die Datenannahme- und Aufbereitungsstelle am ZEGV bereitet die vom FDZ Gesundheit erhaltenen Daten auf. Eine verschlüsselte Version der so aufbereiteten Daten übermittelt die Datenannahme- und Aufbereitungsstelle an die NUKLEUS-Datenhaltung, jedoch ohne den Schlüssel zur Entschlüsselung dieser Daten. Nach Schritt (10) werden die unverschlüsselten Daten in der Datenannahme- und Aufbereitungsstelle gelöscht. Für die Erstellung projektspezifischer Datenauszüge werden die verschlüsselten Daten anlassbezogen aus der NUKLEUS Datenhaltung an die Datenannahme- und Aufbereitungsstelle übermittelt und dort mit dem vorliegenden Schlüssel entschlüsselt.
(9)	NUKLEUS Transferstelle	Nachdem die NAPKON-Daten in der NUKLEUS Transferstelle auf den projektspezifischen Datensatz von NAPKON-Studienteilnehmenden mit Einwilligung sowie benötigten Analyse-Variablen eingeschränkt wurden (ProjDAT1), wird das NAPKON-Pseudonym durch das projektspezifische Pseudonym (TanspPIDx) ersetzt und dieser Datensatz (ProjDAT1 _{TanspPIDx}) von der NUKLEUS Transferstelle an die Analysestelle übergeben. Für jedes Forschungsprojekt wird ein neues projektspezifisches Pseudonym erstellt.
(10)	Datenannahme- und Aufbereitungsstelle	Die Datenannahme- und Aufbereitungsstelle ersetzt das Transport-Pseudonym durch das projektspezifische Pseudonym und erstellt aus den MDAT2 einen projektspezifischen Datenauszug (ProjDAT2 _{TanspPIDx}) und stellt diesen an die Auswertestelle zur Verfügung. Die Auswertestelle kann die beiden projektspezifischen Datensätze ProjDAT1 _{TanspPIDx} und ProjDAT2 _{TanspPIDx} anhand des projektspezifischen Pseudonyms TranspPIDx verknüpfen und gemeinsam analysieren.

Die NUKLEUS- und NAPKON-Infrastruktur ist bereits etabliert und könnte mit der Implementierung der in Abbildung 2 dargestellten technischen Prozesse (nahezu) umgehend starten. Die Vertrauensstelle am RKI und das FDZ Gesundheit sind ebenfalls bereits als Strukturen eingerichtet und erste Test von Datenflüssen sind bereits erfolgt. Der Produktivbetrieb des FDZ Gesundheit ist jedoch erst für 2023 geplant. Wir würden vorab einen zeitnahen Probelauf des Datenflusses aus Abbildung 2 inklusive einer beispielhaften Auswertung eines Use Case aus dem unter Kapitel 1.3 beschriebenen Konzeptes auf Basis der NAPKON POP-Kohorte mit vorliegender KVNR empfehlen. So können mögliche technische und administrative Stolpersteine vor Etablierung des Routinebetriebs identifiziert und ausgeräumt werden. Im Routinebetrieb muss die Datenlieferung aus dem FDZ Gesundheit regelmäßig erfolgen (wenn entweder neue Daten im FDZ Gesundheit vorliegen oder die Kohorte der

NAPKON-Studienteilnehmenden mit Einwilligung in die Datenverknüpfung sich nachhaltig geändert hat). Der zugrundeliegende Datenfluss sollte dabei idealerweise weitgehend automatisiert sein, um den Aufwand für die beteiligten Stellen geringstmöglich zu halten.

Die beschriebene Infrastruktur kann eine Blaupause für andere Forschungsprojekte sein. Je nach Anforderung können die Strukturen angepasst werden.

4.3 Weitere Empfehlungen

Empfehlungen zum konsentbasierten Linkage über Kooperation mit Krankenkassen

Es bestehen weiterhin Szenarien, in denen für ein Linkage auf Basis einer informierten Einwilligung (*informed consent*) die direkte Beteiligung von Krankenkassen erforderlich ist. Dies sind u. a.:

- Projekte, die einen größeren Datenumfang benötigen als am FDZ vorliegend (z. B. Selektivverträge, Pflegedaten)
- Retrospektive Fragestellungen, die über den zunächst im FDZ angebotenen Zeitraum hinausgehen
- Dringende, zeitkritische Analysen (z. B. im Rahmen Evidenzgenerierung bei Pandemie oder Krise)
- Interventionsstudien, wie u. a. Projekte im Bereich der „Neuen Versorgungsformen“ des Innovationsfonds
- Je nach Weiterentwicklung des FDZ Gesundheit Studien mit Record-Linkage

Aufgrund dieses weiteren Bedarfs sollten im Rahmen des Gesundheitsdatennutzungsgesetzes auch Regelungen zur forschungsfreundlichen Umsetzung von GKV-Datenanalysen durch Forschende in direkter Kooperation mit Krankenkassen getroffen werden. Wenn Probanden in das Linkage mit „Krankenkassendaten“ oder „GKV-Routinedaten“ einwilligen, so beinhaltet diese Einwilligung die Nutzung der genannten Daten über die Krankenkasse direkt oder auch über das FDZ Gesundheit. Insofern sollten im Gesundheitsdatennutzungsgesetz beide Nutzungswege von GKV-Routinedaten für Linkagestudien berücksichtigt werden.

Momentan fehlt ein strukturierter, transparenter Entscheidungsprozess für die Bereitstellung von GKV-Daten durch die Krankenkassen. Es fehlt auch ein Gremium, an das kassenübergreifende Anfragen gestellt werden können. Die momentane Praxis birgt die Gefahr von Interessenkonflikten und resultierender selektiver Berichterstattung sowie Publication Bias. Ethisch sind nach der Expertinnen- und Experteneinschätzung die Krankenkassen ihren Versicherten verpflichtet. Die wissenschaftliche Nutzung der Kassendaten sollte als Nutzen für das Gemeinwohl angesehen werden. Sofern mit Auftrag des Versicherten per Einwilligung die Datenübermittlung gewünscht ist, oder "ein hohes

Gemeinwohlnutzenpotential" besteht, liegen starke Argumente für eine Datenübermittlung durch Krankenkassen vor, gegenüber denen das Eigeninteresse von Krankenkassen zurückstehen sollte.

Um Forschung planbar und umsetzbar zu gestalten und um den Entscheidungsprozess der Krankenkassen transparent zu machen, ist ein strukturierter Prozess notwendig. Dieser ist durch eine entsprechende Governance-Struktur und einen transparenten Bewertungsprozess sicherzustellen.

- Es sollten klare Priorisierungskriterien festgelegt werden, anhand derer die Entscheidungen zur Datenbereitstellung getroffen werden (z. B. Qualifikation und Vorerfahrung der Antragsteller, Relevanz der Forschungsfrage, möglicher Impact zur Verbesserung der Gesundheitsversorgung inkl. Prävention, Angemessenheit der Methoden, Datenschutzkonzept der anfordernden Einrichtung)
- Die Entscheidungen sollten nicht mehr von einzelnen KK getroffen werden. Es sollte ein Gremium (KK-Vertreter, Patientenvertreter, Kliniker, Forschende) eingerichtet werden, an das Anträge für ein Linkage mit direkter Beteiligung von Krankenkassen gestellt werden können und das über die Umsetzung entscheidet.
- Die Krankenkassen sollten die anfallenden Kosten für die Datenbereitstellung erstattet bekommen (können). Hier sind entsprechende gesetzliche Regelungen notwendig (Aufnahme der Bereitstellung von Forschungsdaten in den Katalog der Aufgaben der GKV).
- Bei Entscheidung zur Projektumsetzung sollte das Forschungsvorhaben weiterhin der für das Forschungsinstitut zuständigen Ethikkommission vorgelegt und nur bei positivem Votum der Ethikkommission begonnen werden.
- Bei Nutzung von GKV-Daten auf Basis einer informierten Einwilligung und Vorliegen eines zustimmenden Votums der zuständigen Ethikkommission sollte zukünftig kein Antrag nach 75 SGB X notwendig sein.

Allgemeine Empfehlungen zur wissenschaftlichen Nutzung von GKV-Routinedatenanalysen (gleichermaßen über Krankenkassen und FDZ Gesundheit)

Eine entscheidende Barriere für die wissenschaftlich begründete Weiterentwicklung und Verbesserung der Gesundheitsversorgung stellen die im europäischen Vergleich sehr strenge Auslegung der Datenschutzvorgaben in Deutschland dar. In der derzeitigen Praxis dominieren in Deutschland bei Fragen zur Nutzung von Gesundheitsdaten für die wissenschaftliche Forschung die theoretischen Risiken die Diskussionen. Die Chancen und der zu erwartende gesellschaftliche Nutzen der Forschung mit GKV-Routinedaten fällt dagegen kaum ins Gewicht [45]. Um hier eine angemessene Abwägung von Nutzen- und Schadenspotenzialen bei der regulatorischen Bewertung von Forschungsvorhaben vorzunehmen, sollten klar abgrenzbare Kategorien sowohl für Datenschutzrisiken, als auch für Nutzenchancen entwickelt werden, die

dann bei der Abwägung zur Bewilligung oder Ablehnung von Forschungsvorhaben heranzuziehen sind. Patientenvertretende befürworten die wissenschaftliche Nutzung von GKV-Routinedaten, Krankenkassen berichten, dass ihre Versicherten die Daten „in der Regel bereitwillig für die Forschung teilen“ und die klinischen Fachvertreter erwarten von der wissenschaftlichen Nutzung der Daten unmittelbare Vorteile für die Patientenversorgung. Obwohl es in Deutschland keine bekannten Fälle von Missbrauch der Daten im Forschungskontext gibt, macht der Datenschutz viele wichtige Projekte schwierig umsetzbar, wenn nicht gar unmöglich. Bzgl. Datenschutz ist es wesentlich, zwischen theoretischer Separierbarkeit und praktischer Identifizierbarkeit zu unterscheiden. Nicht jede Variablenkombination, die selten vorkommt, verursacht automatisch eine Identifizierbarkeit. Dieser Sachverhalt muss unbedingt bei der Datenschutzbewertung berücksichtigt werden. Angemessene Regelungen zur Vermeidung und Ahndung von Wissenschaftlichem Fehlverhalten (Scientific Misconduct) sind vorhanden. Hinzu tritt eine in Deutschland traditionell stark ausgeprägte Datenschutzkultur in der medizinischen und gesundheitswissenschaftlichen Forschenden-Community.

Momentan können ausschließlich die Daten von GKV-Versicherten für Forschungszwecke nach § 303a ff. SGB V genutzt werden. In dem Ausschluss der privat Versicherten sehen die Gutachter - im Einklang mit den befragten Patientenvertretenden ein Gerechtigkeitsproblem. Das selektive Fehlen der Daten der PKV-Versicherten kann auch je nach Fragestellung zu einer Beeinträchtigung der Qualität der Forschungsergebnisse führen. Im Rahmen des Gesundheitsdatennutzungsgesetzes sollte daher perspektivisch die Einbeziehung von Versorgungsdaten der PKV vorgesehen werden.

Bei GKV-Routinedaten werden Datennutzungen momentan in der Regel mehrfach redundant geprüft. Bei Datennutzung in Kooperation mit einer Krankenkasse werden Fragen zum Datenschutz derzeit typischerweise durch den Datenschutzbeauftragten der Krankenkasse und des Forschungsinstitutes, die Ethikkommission(en), sowie die Aufsichtsbehörde der Krankenkasse geprüft. Wenn mehrere Ethikkommissionen beteiligt sind, so kommen diese nicht immer zu einheitlichen Bewertungen und Empfehlungen (wie etwa in dem Fallbeispiel NAPKON). Für keinen Schritt dieser Prüfkaskade gelten verbindliche Fristen, was Forschung mit GKV-Routinedaten deutlich schlechter planbar macht als etwa klinische Forschung oder Arzneimittelforschung. Es sollten daher für die Ethikkommission wie auch für die Institutionen, die Anträge nach §75 SGB X prüfen, verbindliche Fristen eingeführt werden. Für die Ethikkommissionen sollte weiterhin das pragmatische Konzept der "federführenden Ethikkommission" aus dem Bereich der Studien nach Arzneimittelgesetz auch für GKV-Routinedaten übernommen werden. Nach diesem Konzept ist stets eine Ethikkommission federführend und erstellt in Abstimmung mit anderen "beteiligten" Ethikkommissionen ein gemeinsames Votum innerhalb einer definierten Frist. Diese Maßnahmen könnten den Prüfprozess beschleunigen und verschlanken, ohne Abstriche bei der Qualität der ethischen Bewertung zu verursachen.

Weiter erschwert wird die Situation durch einen teilweise widersprüchlichen Flickenteppich bei den gesetzlichen Erlaubnistatbeständen zur Nutzung von GKV-Routinedaten zu Forschungszwecken. Diesen bestehenden Flickenteppich zu beseitigen, muss ein vordringliches Ziel des Gesundheitsdatennutzungsgesetzes sein. Auch eine Harmonisierung des Sozialdatenschutzes in Deutschland mit der DSGVO sollte in diesem Zusammenhang hergestellt werden. Die momentane Praxis des Datenschutzes und die teilweise unklaren, teilweise widersprüchlichen Regelungen zwischen den Sozialgesetzbüchern werden den Erfordernissen der medizinischen Forschung nicht gerecht, sorgen für Rechtsunsicherheit, und verhindern die evidenzgeleitete Verbesserung der Gesundheitsversorgung. Dieser Zustand gefährdet Gesundheit und Lebensqualität vieler Patientinnen und Patienten und letztlich der Bevölkerung. Wir empfehlen eine bereichsübergreifende gesetzliche Regelung, die die Verarbeitung sämtlicher Behandlungsdaten zu Forschungszwecken grundsätzlich erlaubt. Entsprechende adäquate Governance-Strukturen seitens der Forschenden sind zu etablieren. Hier dienen die NAKO Gesundheitsstudie und die NUM-Infrastruktur als mögliche Blaupausen.

Eine für Forschungsprojekte sinnvolle Hilfestellung ist auch die Bereitstellung von Beispielen zur effektiven Gestaltung von Studienunterlagen mit bereits vorgeprüften Textbausteinen und Formulierungsvorschlägen an einer zentralen Stelle. Sinnvoll wäre auch eine Pflege dieser Unterlagen und ggfs. das Angebot einer Beratungsleistung. Als zentrale Stellen könnten Fachgesellschaften, wie beispielsweise die Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der DGSMP und DGEpi, das DNVF (hier insbesondere die AG Datenlinkage) oder das FDZ Gesundheit fungieren.

Für wissenschaftlich angemessene GKV-Routinedatenanalysen ist eine hohe Expertise und Erfahrung im Umgang mit GKV-Routinedaten erforderlich. Dies beinhaltet neben allgemeiner Methodenkompetenz und praktischer Erfahrung in der Epidemiologie, Versorgungsforschung und Biometrie spezifische Expertise zu den Verfahren der Guten Praxis Sekundärdatenanalyse [42], eine vertiefte Kenntnis der fachlichen Gepflogenheiten und ggf. Besonderheiten in der Dokumentation von Versorgungsprozessen und –outcomes im Forschungsgebiet, sowie eine performante und sichere IT-Umgebung.

Momentan ist der Kreis der Forschenden mit ausgewiesener Expertise im Umgang mit GKV-Daten begrenzt. Um einer breiteren wissenschaftlichen Community den Zugang zu den GKV-Daten zu ermöglichen und gleichzeitig eine angemessene Qualifikation der Forschenden und damit Qualität der Forschung sicherzustellen, sollte eine Schulung angeboten und ein entsprechendes Zertifikat vergeben werden, das mit dem Antrag auf GKV-Routinedatennutzung vorzulegen ist. Eine Orientierung bietet hier das GCP-Zertifikat, das Voraussetzung zur Mitwirkung an einer Studie nach AMG ist.

Zur Nachwuchsförderung und Schaffung einer breiteren Expertise für GKV-Routinedatenanalysen empfehlen wir darüber hinaus auch tiefergehende

Weiterbildungsmaßnahmen, wie z. B. ein Curriculum GKV-Routinedatenanalyse. Vorbild könnte hier das Curriculum für Registerstudien in der Onkologie (REGIBA) sein [73], das momentan unter Förderung des BMG durch das DNVF entwickelt wird.

Momentan ist eine wissenschaftliche Nutzung von GKV-Routinedaten weitgehend von der Entscheidung einzelner Krankenkassen abhängig. Krankenkassen bewerten die Relevanz des Projekts aus ihrer jeweils eigenen Perspektive, woraus ein Interessenkonflikt resultieren kann. Es besteht weiterhin keine Transparenz, welche Forschung mit GKV-Routinedaten betrieben wird. Aus der klinischen Forschung wissen wir, dass prospektive Studienregistrierung eine wirksame und gut umsetzbare Maßnahme ist, um selektive Berichterstattung und Publication Bias entgegen zu wirken. Es sollte daher eine prospektive Registrierung aller wissenschaftlicher GKV-Routinedatenanalysen verbindlich eingeführt werden. Dies kann in einem spezifischen Register für GKV-Routinedatenanalysen oder in einem der etablierten Studienregister umgesetzt werden.

Momentan fehlt eine Transparenz über das aktuelle Versorgungsgeschehen. Dies betrifft insbesondere die regionale Betrachtungsebene, die für die Bewertung der Angemessenheit der Gesundheitsversorgung für die Bürger besonders relevant ist. Das Forschungsdatennutzungsgesetz sollte die Möglichkeit der Nutzung der Daten des FDZ Gesundheit für das (regionale) Versorgungsmonitoring, etwa von örtlich-zeitlichen Erkrankungsspitzen bei infektiösen Erkrankungen, Eingriffszahlen elektiver Operationen oder das Verordnungsverhalten von Schmerzmitteln und von Antibiotika vorsehen. In diesem Fall sollten regionale Netzwerke die Möglichkeit bekommen, auf Antrag kontinuierlich (z. B. alle 4 Wochen) aktuelle Daten zu ausgewählten Schlüsselindikatoren vom FDZ Gesundheit zu erhalten. Für eine solche Nutzung wie auch generell sollte das Zeitintervall zwischen Datenentstehung und Datennutzbarkeit durch das FDZ Gesundheit minimiert werden.

Weitere Anforderungen entstehen im Zuge der Umsetzung des European Health Data Space (EHDS). Wichtige Ziele des EHDS sind die Erhöhung der Kontrolle der Bürgerinnen und Bürger über ihre eigenen Gesundheitsdaten, und die Förderung der Nutzung von Gesundheitsdaten für eine bessere medizinische Versorgung, für Forschung, Innovation und Politikgestaltung. EHDS soll allen EU-Ländern ermöglichen, das Potenzial von Austausch, Nutzung und Weiterverwendung von Gesundheitsdaten unter gesicherten Bedingungen auszuschöpfen. Das FDZ Gesundheit könnte in diesem Kontext als nationale Zugangsstelle für elektronische Gesundheitsdaten ausgebaut werden, da hier ein bedeutsamer bundesweiter Datensatz vorliegt und die personellen, technischen und datenschutzrechtlichen Voraussetzungen für eine zentrale Rolle im EHDS im Grundsatz vorliegen.

5 Zusammenfassung

Versorgungsnahe Daten (VeDa) sind die wesentliche Ressource für die wissenschaftlich fundierte Beschreibung, Evaluation und Weiterentwicklung der Gesundheitsversorgung und des Gesundheitssystems in Deutschland. Eine ganz wesentliche Quelle von VeDa sind Abrechnungsdaten der Gesetzlichen Krankenversicherung (GKV-Routinedaten). Das Forschungsdatenzentrum Gesundheit (FDZ Gesundheit) am Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) hat für die Gesundheitsforschung mit VeDa eine überragende Bedeutung, weil es Längsschnittdaten von allen mehr als 70 Millionen GKV-Versicherten in Deutschland sektorübergreifend zur ambulanten und stationären medizinischen Versorgung, sowie zur Arzneimittelversorgung verfügbar machen wird. Trotz der hohen Repräsentativität und Vollständigkeit der dokumentierten Diagnosen, Prozeduren und Arzneimittelverordnungen fehlen jedoch in den GKV-Routinedaten häufig wichtige zusätzliche Informationen zur Beantwortung von wichtigen Forschungsfragestellungen, etwa zu Patienten-berichteten Ergebnissen, zu soziodemografischen Faktoren, zur Diagnostik und Bildgebung, zu Laborergebnissen und Biomarkern sowie zu Patientenpräferenzen. Für die qualitativ hochwertige wissenschaftliche Beantwortung von versorgungsrelevanten Fragestellungen müssen die GKV-Routinedaten daher häufig mit anderen VeDa (Registerdaten, Klinikdaten) oder/und mit Befragungsdaten und Daten aus Kohortenstudien individuell verknüpft werden (Data Linkage).

Das vorliegende Gutachten analysiert zunächst den derzeitigen Status quo der Umsetzung von Studien mit Linkage von GKV-Routinedaten mit Forschungsdaten basierend auf der informierten Einwilligung von Studienteilnehmenden. Dies wird exemplarisch am Beispiel der NAPKON-Kohortenstudie des Netzwerks Universitätsmedizin (NUM) betrachtet, welche zur Untersuchung von Verläufen einer Coronavirus-Erkrankung (COVID-19) und von Risikofaktoren für eine Long-COVID Erkrankung primäre Forschungsdaten mit GKV-Routinedaten verknüpfen möchte. Unter Einbeziehung der Perspektive von Patientenvertretenden, Krankenkassenvertretenden, Expertinnen und Experten für Ethik, Forschenden sowie von Klinikerinnen und Klinikern erarbeitet das Gutachten darauf aufbauend eine Modelllösung für die Verknüpfung der im FDZ Gesundheit zukünftig vorgehaltenen GKV-Routinedaten mit externen Forschungsdaten und geht insbesondere auf technisch-organisatorische Voraussetzungen ein. Ausgehend von dem Fallbeispiel NAPKON, eigenen Erfahrungen der Gutachter und den Ergebnissen der Expertenbefragungen zeigt das Gutachten weiteren gesetzlichen Regelungsbedarf zur Forschung mit GKV-Routinedaten auf und gibt entsprechende Empfehlungen für die Berücksichtigung in dem von der Bundesregierung im Koalitionsvertrag angekündigten Gesundheitsdatennutzungsgesetz.

Die gegenwärtigen Rahmenbedingungen für die wissenschaftliche Nutzung von Behandlungsdaten in Deutschland sind unbefriedigend. Das große Potenzial, die Qualität,

Sicherheit und Effizienz der medizinischen Versorgung durch die Versorgungsforschung zu verbessern sowie geeignete Public-Health-Maßnahmen zu planen und umzusetzen wird bisher bei Weitem nicht genutzt. Es bestehen momentan erhebliche datenschutzrechtliche, regulatorische und organisatorische Hürden und in der Konsequenz eine fehlende Verlässlichkeit und Planbarkeit bzgl. der Möglichkeit der Verknüpfung von GKV-Routinedaten mit Forschungsdaten. In der derzeitigen Praxis dominieren in Deutschland bei Fragen zur Nutzung von Gesundheitsdaten für die wissenschaftliche Forschung die theoretischen Risiken die Diskussionen und datenschutzrechtlichen Entscheidungen. Die Chancen und der zu erwartende gesellschaftliche Nutzen der Forschung mit GKV-Routinedaten fallen dagegen kaum ins Gewicht.

Um ein angemessenes Abwägen von Nutz- und Schadenspotenzialen bei der regulatorischen Bewertung von Forschungsvorhaben zu ermöglichen, sollten klar abgrenzbare Kriterien sowohl für Datenschutzrisiken, als auch für Nutzenchancen entwickelt werden, die dann transparent bei der Abwägung zur Bewilligung oder Ablehnung von Forschungsvorhaben heranzuziehen sind. Patientenvertreter befürworten die wissenschaftliche Nutzung von GKV-Routinedaten, Krankenkassen berichten, dass ihre Versicherten die Daten „in der Regel bereitwillig für die Forschung teilen“ und die klinischen Fachvertreter erwarten von der wissenschaftlichen Nutzung der Daten einen unmittelbaren Nutzen für die Patientenversorgung und die medizinische Qualität. Obwohl es in Deutschland keine bekannten Fälle von Missbrauch der Daten im Forschungskontext gibt, macht der Datenschutz viele wichtige Projekte schwierig umsetzbar, wenn nicht gar unmöglich. Bzgl. Datenschutz ist es wesentlich, zwischen theoretischer Separierbarkeit und praktischer Identifizierbarkeit zu unterscheiden. Wenn eine Variablenkombination selten vorkommt, verursacht dies nicht automatisch eine Identifizierbarkeit. Die Identifizierung einer bestimmten Person in einem pseudonymisierten Datensatz erfordert einen hohen Aufwand und in aller Regel kriminelle Energie. Dieser Sachverhalt muss unbedingt bei der Datenschutzbewertung berücksichtigt werden. Angemessene Regelungen zur Vermeidung und Ahndung von Scientific Misconduct existieren seit Jahren und sind konsequent anzuwenden. In der gesundheitswissenschaftlichen Forschungs-Community existiert eine hohe Datenschutzorientierung. Neben den klaren Sanktionsmöglichkeiten ist es vor allem diese ausgeprägte Kultur der Selbstkontrolle, die Missbrauchsrisiken minimiert hat.

Auch bei informierter Einwilligung von Studienteilnehmende in die Verknüpfung der über sie erhobenen Forschungsdaten mit ihren bei ihrer Krankenkasse gespeicherten GKV-Routinedaten ist es im Beispiel der NAPKON-Kohortenstudie bisher nicht gelungen, GKV-Daten tatsächlich für die wissenschaftliche Nutzung zu erhalten. Gründe sind redundante, nicht standardisierte, restriktive Prüfungen des Datenschutzes durch unterschiedliche Institutionen (Datenschützer der Einrichtungen, Ethikkommissionen, Aufsichtsbehörden der Krankenkassen) mit häufig widersprüchlichen Ergebnissen sowie eine fehlende Verpflichtung

der Krankenkassen, dem Willen ihrer Versicherten zur Verfügbarmachung der GKV-Routinedaten zu entsprechen.

Aktuell können Forschungsprojekte mit Linkage von sektorenübergreifenden GKV-Routinedaten nur mit aktiver Unterstützung durch Krankenkassen durchgeführt werden. Größere Krankenkassen erhalten sehr viele Anfragen zur Forschungsdatennutzung, denen sie häufig nicht entsprechen können oder nach eigener Priorisierung nicht entsprechen wollen. Es fehlen transparente Entscheidungsprozesse und kassenübergreifende Gremien zur Koordination der Forschungsanfragen. Die momentane Praxis birgt die Gefahr von Interessenkonflikten und resultierender selektiver Berichterstattung/Publication Bias. Krankenkassen und ihre wissenschaftlichen Institute haben vielfach eine hohe Expertise in der Versorgungs-, Qualitäts- und Patientensicherheitsforschung mit GKV-Routinedaten. Sie sind häufig nicht nur für die Datenbereitstellung zuständig, sondern auch für die Planung und Umsetzung von Forschungsprojekten. Dennoch und gerade deshalb sollte eine klarere Transparenz darüber hergestellt werden, wer was mit GKV-Routinedaten beforcht. Wirksame Maßnahmen wie prospektive Studienregistrierung sowie Berichts- und Qualitätsstandards sind aus der Klinischen Forschung bekannt und sollten für die Versorgungsforschung mit GKV-Routinedaten ebenfalls umgesetzt werden. Da auch nach erfolgreicher Implementierung des FDZ Gesundheit weiterhin mehrere Szenarien bestehen werden, bei denen die Krankenkassen für die Datenbereitstellung direkt beteiligt werden müssen (z. B. Interventionsstudien, Evaluation Neuer Versorgungsformen des Innovationsfonds, Versorgungsmanagement, Studien, die GKV-Daten benötigen, die momentan nicht für das FDZ Gesundheit vorgesehen sind), schlägt das Gutachten entsprechende Regelungen zur Erhöhung der Planbarkeit von Forschungsprojekten und der Transparenz von Entscheidungsprozessen für oder gegen die Unterstützung von Forschungsprojekten durch die Krankenkassen vor.

Bei Nutzung von GKV-Daten auf Basis einer informierten Einwilligung und Vorliegen eines zustimmenden Votums der zuständigen Ethikkommission sollte zukünftig kein Antrag nach § 75 SGB X notwendig sein. Hintergrund für diese Empfehlung ist, dass alle Voraussetzungen für die Sozialdatennutzung bereits durch die zuständige Ethikkommission geprüft werden und durch die Mehrfachprüfung kein Datenschutznutzen, aber Aufwände auf Forschungsseite produziert werden. Das FDZ Gesundheit hat das Potenzial, entscheidend zum Ausbau der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland beizutragen. Dies ist ethisch und politisch geboten, um die Qualität, Sicherheit und Wirtschaftlichkeit der Gesundheitsversorgung evidenzgeleitet zu erhöhen, neue medizinische Anwendungen zu entwickeln und damit die Gesundheit und Lebensqualität der Bevölkerung zu verbessern. Um dieses Potenzial zu heben und den politischen Willen einer breiten und qualitativ hochwertigen wissenschaftlichen Nutzung von GKV-Routinedaten zu erfüllen, muss das FDZ Gesundheit für Wissenschaftlerinnen und Wissenschaftler praktikable Lösungen vorsehen. Dazu zählt neben der Verfügbarmachung möglichst aktueller Daten u. a. die Möglichkeit, per personalisiertem Remotezugang pseudonymisierte Einzeldatensätze vom FDZ Gesundheit

analysieren zu können. Um den wissenschaftlichen Anforderungen der Versorgungsforschung mit GKV-Routinedaten zu entsprechen, sollten Forschende nach Bewilligung von Nutzungsanträgen diesen Remotezugang zu einem möglichst breiten Datensatz erhalten. Dies ist notwendig, weil bei GKV-Routinedatenanalysen – anders als etwa bei Klinischen Studien – ohne genaue Kenntnis des Datenkörpers nicht alle Analyseschritte abschließend geplant werden können. Synthetische Daten oder „verrauschte“ Daten bzw. Anwendung von differential privacy oder verteiltes Rechnen sind für die Planung und Durchführung von GKV-Routinedatenanalysen nach aktuellem Kenntnisstand nicht geeignet.

Einige für die Gesundheitsversorgung äußerst relevante Fragestellungen können nur mit einem vollständigen, krankenkassenübergreifenden Datensatz aller Versicherter in der notwendigen wissenschaftlichen Qualität bearbeitet werden. Hierzu zählen beispielsweise Fragestellungen der Qualitätssicherung, der Patientensicherheitsforschung, der Arzneimittelsicherheit, des kleinräumigen Versorgungsmonitorings, zu seltenen Erkrankungen sowie zur Leitliniengerechtigkeit der Versorgung in spezifischen Patientengruppen. Die Einrichtung des FDZ Gesundheit ermöglicht die Forschung in diesen wichtigen Arbeitsfeldern. Bereits einzelne fehlende Daten können diese Ergebnisse verzerren oder eine Frage gänzlich unbeantwortbar machen. Aus diesem Grund sollte die Übermittlung der GKV-Daten an das FDZ Gesundheit ohne individuelle Widerspruchsmöglichkeit erfolgen.

Aktuell ist die Datenlieferung von den Krankenkassen an das FDZ im jährlichen Turnus festgeschrieben. Je nach Forschungsinhalt oder Dringlichkeit von datenbasierten Entscheidungen (bspw. Pandemische Gefährdungslage) sollten Lieferungen im kürzeren Zeitabständen möglich sein. Die COVID-19 Pandemie hat gezeigt, dass GKV-Routinedaten für gesamtgesellschaftlich hochrelevante Fragestellungen eine ganz wesentliche, wenn nicht die einzige Datenbasis für evidenzbasierte Entscheidungen darstellen, dass GKV-Routinedaten aber in Deutschland im internationalen Vergleich sehr viel langsamer und unvollständiger oder nicht direkt verlinkbar (z. B. Impfdatenprojekt RiCO [74]) verfügbar waren. Die Verlinkbarkeit insbesondere bei neuen Datenerhebungen sollte aus gesetzgeberischer Sicht immer mitgedacht werden. Vor dem Hintergrund einer notwendigen besseren Vorbereitung auf zukünftige Pandemien und Krisen (Pandemic Preparedness) und für die Sicherstellung einer internationalen Wettbewerbsfähigkeit der Versorgungsforschung sollte eine sektorenbezogene Beschleunigung der Liefertermine an das FDZ Gesundheit erfolgen. Stationäre Aufenthalte können bereits wenige Tage nach Entlassdatum übermittelt und Verordnungsdaten innerhalb eines Monats verfügbar gemacht werden. Auch die quartalsmäßige Abrechnung der ambulanten ärztlichen Daten über die Kassenärztlichen Vereinigungen (KV) an die Krankenkassen sollten unter dem Gesichtspunkt der schnelleren Nutzarmachung auf den Prüfstand gestellt werden. So könnten in einer Art "fast lane" Verfahren kritische Ereignisse tagesgenau berichtet werden, beispielsweise der Anstieg von Atemwegserkrankungen in der ambulanten Versorgung, als Frühwarnsystem, noch bevor es

zu einer Häufung von stationären Aufnahmen kommt. Dieses Verfahren sollte flexibel auf die jeweilige (heterogene) Gefährdungslage reagieren können.

Momentan ist der Kreis der Forschenden mit ausgewiesener Expertise im Umgang mit GKV-Daten in Deutschland begrenzt. Um einerseits einen breiten wissenschaftlichen Zugang zu den Daten des FDZ Gesundheit zu ermöglichen und gleichzeitig eine hinreichend hohe Qualität der Forschungsarbeiten mit den Daten sicherzustellen, sollte ein Zertifikat entwickelt werden, das Voraussetzung zur Datennutzung wird. Eine mögliche Referenz bildet das für die Beteiligung an Arzneimittelstudien notwendige GCP (Good Clinical Practice)-Zertifikat. Zusätzlich zu einem Zertifikat „Gute Praxis Sekundärdatenanalyse und Linkage“ sollte zur Nachwuchsförderung ein Curriculum für die wissenschaftliche Nutzung und das Linkage von GKV-Routinedaten entwickelt werden.

Momentan können ausschließlich die Daten von GKV-Versicherten für Forschungszwecke nach § 303a ff. SGB V genutzt werden. Der Ausschluss der privat Versicherten stellt nach Überzeugung der Gutachter und der befragten Patientenvertretungen ein Gerechtigkeitsproblem dar. Das selektive Fehlen der Daten der PKV-Versicherten kann auch je nach Fragestellung zu einer Beeinträchtigung der Qualität der Forschungsergebnisse führen. Der Gesetzgeber sollte daher auch die perspektivische Einbeziehung von Versorgungsdaten der PKV in das FDZ Gesundheit vorsehen. Hierbei kann auf erste Projekte, die Versorgungsdaten innerhalb der PKV zu nutzen, aufgesetzt werden.

Das größte wissenschaftliche Potenzial zur Verbesserung der Gesundheitsversorgung und Gesundheit der Bevölkerung liegt im Linkage von GKV-Routinedaten mit anderen Daten. Dies ist heute schon Goldstandard etwa in der NAKO Gesundheitsstudie oder beim Broad Consent der Medizininformatik Initiative. Ein Linkage ist i. d. R. notwendig, um kausale Schlüsse zu ziehen. Wir sehen daher die Linkagemöglichkeit der GKV-Daten mit anderen Datenquellen als eine entscheidende Voraussetzung für den Erfolg des FDZ Gesundheit. Wie die NAKO Gesundheitsstudie zeigt, ist die Akzeptanz der Studienteilnehmenden zur Nutzung und Linkage von GKV-Routinedaten mit Primärforschungsdaten groß. Grundsätzlich sollte das Linkage wie im Fallbeispiel dieses Gutachtens auf Basis einer informierten Einwilligung (informed consent) erfolgen. In der Rechtsgrundlage des FDZ Gesundheit (§303 a bis f SGB V) sowie der zugehörigen Datentransparenzverordnung findet sich derzeit kein Hinweis auf ein konsentbasiertes Datenlinkage, unabhängig davon, ob das Linkage beim Forscher (nach Export der Daten durch das FDZ Gesundheit) oder im FDZ stattfindet (mit Übermittlung der klinischen Primärdaten aus der anfragenden Studie an das FDZ). Eine entsprechende Anpassung der Datentransparenzverordnung wird daher dringend empfohlen.

Im Rahmen größerer prospektiver Kohortenstudien und Register mit konsentbasiertem Datenlinkage ist für die Forschungsinfrastruktur der Studiengruppe eine Datenannahme- und Aufbereitungsstelle, eine Datentreuhänderstelle sowie ein multiperspektivisches Use &

Access Board vorzusehen. Nach Genehmigung des Forschungsvorhabens und Vorliegen eines entsprechenden Ethikvotums erfolgt die Lieferung der GKV-Routinedaten vom FDZ Gesundheit an die Forschungsinfrastruktur der Studiengruppe, die das Datenlinkage dann fragestellungsbezogen umsetzt. Diese Vorgehensweise ist bereits gängige Praxis bei der Kooperation von Forschenden mit Krankenkassen und sollte analog bei der zukünftigen Verfügbarmachung der GKV-Routinedaten durch das FDZ Gesundheit umgesetzt werden.

Für das konsentbasierte Linkage sind im FDZ Gesundheit technische, organisatorische und rechtliche Rahmenbedingungen zu schaffen. Die vorliegende Einwilligung inkl. der Zustimmung zum Datenlinkage muss als Grundlage für die Nutzung von Daten des FDZ Gesundheit durch entsprechende Regelungen juristisch belastbar anerkannt werden. Bei Nutzung der Daten von Krankenkassen und Vorliegen einer Einwilligung sollte auf einen Antrag nach §75 SGB X verzichtet oder das Antragsverfahren beschleunigt werden.

Die Vertrauensstelle DaTraV des FDZ Gesundheit muss organisatorisch in die Prozesse für ein konsentbasiertes Linkage einbezogen werden, damit die in Studienteilnehmerlisten bereitgestellte KVNR in das zugehörige PüP transferiert und an das FDZ Gesundheit projektbezogen übermittelt werden. Wenn zusätzlich Studiendaten übermittelt werden, müssen auch diese an das FDZ Gesundheit weitergeleitet und dort zusammen mit den verlinkten Routinedaten bereitgestellt werden. Hierfür sind im FDZ Gesundheit entsprechende logistische Voraussetzungen sowie die notwendige IT-Infrastruktur zu schaffen, um diese Daten nach der Verknüpfung in einer sicheren Umgebung bereitstellen zu können. Bei Anfrage eines Forschungsprojekts auf Übermittlung von GKV-Routinedaten sind entsprechende Prozesse für die Bereitstellung von Einzeldatensätzen der Studienteilnehmenden zu schaffen. Das Datenlinkage zwischen den Datenquellen erfolgt über temporäre Lieferpseudonyme, die zusammen mit der KVNR übermittelt wurden und die eindeutige Zuordnung der Individualdaten ermöglichen. Bei Datenexport sind diese Lieferpseudonyme in den übermittelten Datensatz zu integrieren.

Das Vorhalten eines Versichertenverzeichnisses in einer Vertrauensstelle DaTraV im Sinne eines Identitätsmanagements wird empfohlen. Hierdurch verbessert sich die Qualität und es eröffnen sich weitere Funktionen und Möglichkeiten. Aus den Daten des FDZ Gesundheit vorselektierte Versicherte könnten angeschrieben werden, um an spezifischen Studien teilzunehmen oder es könnten Daten gezielt bei den Versicherten nacherhoben werden. Dies eröffnet insbesondere für seltene Erkrankungen ein enormes Forschungspotential. Auf Basis des Versichertenverzeichnisses können auch Studien- oder Registerdaten ohne Vorliegen der KVNR verlinkt werden, indem die in beiden Datenquellen verfügbaren identifizierenden Angaben des Probanden/der Probandin verlinkt werden. Das wechselseitige Datenlinkage zweier Datenquellen (z. B. Erkrankungsregister mit GKV Gesundheit) ermöglicht die Bearbeitung neuer Fragestellungen, die allein auf Basis der jeweils verfügbaren Einzeldatenquellen nicht möglich wäre. Alternativ zum Aufbau eines bundesweiten

Versichertenverzeichnis in der Vertrauensstelle DaTraV beim RKI können auch bestehende Strukturen und Stellen, z. B. Datenstelle der Rentenversicherung (DSRV), diese Funktion übernehmen.

In besonderen Fällen, wenn Nutzenpotenziale beträchtlich sind, das öffentliche Interesse an der Forschung oder Planung erheblich ist und eine Einwilligung der betroffenen Personen nicht umsetzbar oder unzumutbar ist, sollte das Linkage komplementärer VeDa auch ohne Einwilligung ermöglicht werden. In diesem Fall sollte – unabhängig davon, ob die GKV-Daten vom FDZ Gesundheit oder von Krankenkassen direkt bezogen werden – ein Antrag nach §75 SGB X auf die Übermittlung von Sozialdaten für die Forschung gestellt werden. Ein Beispiel bildet hier das Linkage von GKV-Daten mit Daten der Klinischen Krebsregister (KKR). Das vom Innovationsfonds geförderte Projekt WiZen zur Wirksamkeit der Versorgung in onkologischen Zentren zeigte, dass das Linkage von GKV-Daten mit Daten ausgewählter KKR anhand der pseudonymisierten Krankenversicherungsnummer (KVNR) realisiert werden kann. Ein momentanes Hindernis für die grundsätzliche Ermöglichung eines Linkages bundesweiter KKR-Daten mit GKV-Daten ist die Uneinheitlichkeit der landesgesetzlichen Regelungen in den Krebsregistern. Hier ist dringend eine Harmonisierung notwendig, damit der große Datenschatz, der durch die Verbindung von GKV-Routinedaten und KKR-Daten entsteht, für die onkologische Spitzenforschung, beispielsweise für die Ableitung von Leitlinienempfehlungen genutzt werden kann und damit die medizinische Versorgung von krebserkrankten Menschen in Deutschland verbessert werden kann.

6 Danksagung

Wir danken den folgenden Personen für die Unterstützung bei der Erarbeitung der Gutachtenempfehlungen durch Beteiligung an Interviews, Gruppendiskussionen oder die Beantwortung eines Fragebogens: Prof. Dr. Eva Winkler (Deutsches Krebsforschungszentrum Heidelberg), Prof. Dr. Sabine Salloch (Medizinischen Hochschule Hannover), Prof. Dr. Kai Cornelius (Universität Hamburg), Prof. Dr. Jan Schildmann (Universität Halle), Anja Schramm (AOK Bayern), Prof. Dr. Enno Swart (Universität Magdeburg), Prof. Dr. Falk Hoffmann (Universität Oldenburg), Antke Wolter (DAK), Prof. Dr. Wolfgang Ahrens (Leibniz-Institut für Präventionsforschung und Epidemiologie), Christian Günster (Wissenschaftliches Institut der AOK), Dr. Udo Schneider (Techniker Krankenkasse), Claudia Schulte (BARMER Institut für Gesundheitssystemforschung), Josephine Jacob (Institut für angewandte Gesundheitsforschung Berlin), Thomas Steinbach (Selbsthilfegruppe Prostatakrebs BPS), Corinna Elling-Audersch (Deutsche Rheuma-Liga), Simon Schöning (Selbsthilfegruppe Long COVID Deutschland), Dr. med. Ulrike Holtkamp (Deutsche Leukämie- & Lymphom-Hilfe), Karin Arndt (Sprecherin (Patientenbeirat NCT Partnerstandort Dresden), Prof. R. Thyrian (Deutsches Zentrum für Neurodegenerative Erkrankungen, DZNE), Barbara Mitschdörfer (Bundesverband "Das frühgeborene Kind"), Hedy Kerek-Bodden (Haus der Krebshilfe), Sabine Leitner (Bundesverband "Das frühgeborene Kind"), Wolfgang Krämer (Patientenbeirat NCT Partnerstandort Dresden), Prof. Dr. Jürgen Deckert (Universitätsklinikum Würzburg), Prof. Dr. Marcus Dörr (Universitätsmedizin Greifswald), Dr. Ekaterina Slotina (Universitätsklinikum Jena), Prof. Dr. Sandra Frank (LMU Klinikum München), PD Dr. Antje Freytag (Universitätsklinikum Jena), Timo Fuchs (Universitätsklinikum Regensburg), Prof. Dr. Ildiko Gagyor (Universitätsklinikum Würzburg), Tillmann Görig (Universitätsmedizin Greifswald), Prof. Dr. Stefanie Joos (Universitätsklinik Tübingen), Prof. Dr. Christine Klein (UKSH Lübeck), Prof. Dr. Christina Lemhöfer (Universitätsklinikum Jena), Dr. Robby Markwart (Universitätsklinikum Jena), Prof. Dr. Hans-Joachim Mentzel (Universitätsklinikum Jena), Prof. Dr. Matthias Nauck (Universitätsmedizin Greifswald), Prof. Dr. Ahmed Othman (Universitätsmedizin Mainz), Dr. Silke Schrum (Universitätsklinikum Hamburg-Eppendorf), Prof. Dr. Sophia Stöcklein (LMU Klinikum München), PD Dr. Martin Weigl (LMU Klinikum München), Prof. Dr. Winfried Häuser (Klinikum Saarbrücken), Prof. Dr. Patrick Meybohm (Universitätsklinikum Würzburg), Daniela Fröhlich (Universitätsklinik Tübingen), Prof. Dr. Dr. Maria Cristina Polidori (Universitätsklinikum Köln), Prof. Dr. Peter Heuschmann (Universitätsklinikum Würzburg), Prof. Dr. Sebastian-Edgar Baumeister (Universität Münster) Weiterhin danken die Gutachter Frau Alina Bernhard für die Beratung bzgl. der juristischen Aspekte zur Anwendung des § 75 SGB X im Falle einer informierten Einwilligung zur GKV-Datennutzung durch Teilnehmende an Forschungsprojekten.

7 Literaturverzeichnis

1. Klinkhammer-Schalke M, Kaiser T, Apfelbacher C, Benz S, Dreinhofer KE, Geraedts M, et al. [Manual for Methods and Use of Routine Practice Data for Knowledge Generation]. *Gesundheitswesen*. 2020;82(8-09):716-22. Epub 2020/09/23. doi: 10.1055/a-1237-4011. PubMed PMID: 32961567; PubMed Central PMCID: PMCPMC7661315.
2. Hoffmann F, Kaiser T, Apfelbacher C, Benz S, Bierbaum T, Dreinhofer K, et al. [Routine Practice Data for Evaluating Intervention Effects: Part 2 of the Manual]. *Gesundheitswesen*. 2021;83(6):470-80. Epub 2021/05/22. doi: 10.1055/a-1484-7235. PubMed PMID: 34020493.
3. IQWiG. Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V. Rapid Report A19-43 [Internet]. 2020; (Version 1.1). Available from: https://www.iqwig.de/download/a19-43_versorgungsnaher-daten-zum-zwecke-der-nutzenbewertung_rapid-report_v1-1.pdf.
4. Schons M, Pilgram L, Reese JP, Stecher M, Anton G, Appel KS, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. *Eur J Epidemiol*. 2022;37(8):849-70. Epub 2022/07/30. doi: 10.1007/s10654-022-00896-z. PubMed PMID: 35904671; PubMed Central PMCID: PMCPMC9336157.
5. Peters A, German National Cohort C, Peters A, Greiser KH, Gottlicher S, Ahrens W, et al. Framework and baseline examination of the German National Cohort (NAKO). *Eur J Epidemiol*. 2022;37(10):1107-24. Epub 2022/10/20. doi: 10.1007/s10654-022-00890-5. PubMed PMID: 36260190; PubMed Central PMCID: PMCPMC9581448.
6. Volzke H, Schossow J, Schmidt CO, Jurgens C, Richter A, Werner A, et al. Cohort Profile Update: The Study of Health in Pomerania (SHIP). *Int J Epidemiol*. 2022. Epub 2022/03/30. doi: 10.1093/ije/dyac034. PubMed PMID: 35348705.
7. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort profile: the study of health in Pomerania. *Int J Epidemiol*. 2011;40(2):294-307. Epub 2010/02/20. doi: 10.1093/ije/dyp394. PubMed PMID: 20167617.
8. Hosten N, Bulow R, Volzke H, Domin M, Schmidt CO, Teumer A, et al. SHIP-MR and Radiology: 12 Years of Whole-Body Magnetic Resonance Imaging in a Single Center. *Healthcare (Basel)*. 2021;10(1). Epub 2022/01/22. doi: 10.3390/healthcare10010033. PubMed PMID: 35052197; PubMed Central PMCID: PMCPMC8775435.
9. Lange M, Butschalowsky HG, Jentsch F, Kuhnert R, Schaffrath Rosario A, Schlaud M, et al. [The first KiGGS follow-up (KiGGS Wave 1): study conduct, sample design, and response]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*. 2014;57(7):747-61. Epub 2014/06/22. doi: 10.1007/s00103-014-1973-9. PubMed PMID: 24950824.
10. Lange M, Hoffmann R, Mauz E, Houben R, Gosswald A, Rosario AS, et al. KiGGS Wave 2 longitudinal component - data collection design and developments in the numbers of participants in the KiGGS cohort. *J Health Monit*. 2018;3(1):92-107. Epub 2018/03/15. doi: 10.17886/RKI-GBE-2018-035. PubMed PMID: 35586182; PubMed Central PMCID: PMCPMC8848915.
11. Niemeyer A, Kluge S, Gurisch C, Hoffmann W, Kostuj T, Olbrich K, et al. [Position paper of the German Network for Health Services Research (DNVF) on application-related data

- collection according to Social Code Book V]. *Gesundheitswesen*. 2021;83(4):309-13. Epub 2021/04/09. doi: 10.1055/a-1391-3908. PubMed PMID: 33831973.
12. Jacob J, Walker J, Swart E, Baum F, Rossler M, Tesch F, et al. [Recommendations for the Utilization of Claims Data During a Pandemic: Lessons Learned from the Project EgePan-Unimed of the Netzwerk Universitätsmedizin (NUM)]. *Gesundheitswesen*. 2022. Epub 2022/09/29. doi: 10.1055/a-1915-4526. PubMed PMID: 36170864.
 13. Ihle P, Schneider K, Heß S. Das Forschungsdatenzentrum Gesundheit – Routinedaten der Gesetzlichen Krankenversicherung für die Gesundheits- und Versorgungsforschung. *Gesundheitswesen aktuell* [Internet]. 2022:[32-47 pp.].
 14. Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen. Digitalisierung für Gesundheit Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystems. Gutachten 2021 [Internet]. 2021. Available from: <https://www.svr-gesundheit.de/gutachten/gutachten-2021/>.
 15. Seidler A, Wagner M, Schubert M, Droge P, Pons-Kuhnemann J, Swart E, et al. Myocardial Infarction Risk Due to Aircraft, Road, and Rail Traffic Noise. *Dtsch Arztebl Int*. 2016;113(24):407-14. Epub 2016/07/07. doi: 10.3238/arztebl.2016.0407. PubMed PMID: 27380755; PubMed Central PMCID: PMC4939427.
 16. Hasselhorn HM, Peter R, Rauch A, Schroder H, Swart E, Bender S, et al. Cohort profile: the lidA Cohort Study-a German Cohort Study on Work, Age, Health and Work Participation. *Int J Epidemiol*. 2014;43(6):1736-49. Epub 2014/03/13. doi: 10.1093/ije/dyu021. PubMed PMID: 24618186; PubMed Central PMCID: PMC4276057.
 17. Druschke D, Arnold K, Heinrich L, Reichert J, Rudiger M, Schmitt J. Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects. *Gesundheitswesen*. 2020;82(S 02):S108-S16. Epub 2020/03/21. doi: 10.1055/a-1082-0740. PubMed PMID: 32193879.
 18. Hoffmann W. Nutzung von Gesundheitsdaten in prospektiven Kohortenstudien 2022 [03.12.2022]. Available from: https://www.bfdi.bund.de/SharedDocs/Kurzmeldungen/DE/2022/18_Symposium-Forschungsdaten.html.
 19. Drepper J. Stand der Consent-Nutzung an den Standorten der MII. Internes Papier der AG Consent - TMF eV (15/09/2022),2022.
 20. Horn A, Krist L, Lieb W, Montellano FA, Kohls M, Haas K, et al. Long-term health sequelae and quality of life at least 6 months after infection with SARS-CoV-2: design and rationale of the COVIDOM-study as part of the NAPKON population-based cohort platform (POP). *Infection*. 2021;49(6):1277-87. Epub 2021/10/14. doi: 10.1007/s15010-021-01707-5. PubMed PMID: 34642875; PubMed Central PMCID: PMC8508400.
 21. Netzwerk Universitätsmedizin [12.12.2022]. Available from: <https://www.netzwerk-universitaetsmedizin.de/>.
 22. Weichert T. Datenschutzrechtliche Rahmenbedingungen medizinischer Forschung - Vorgaben der EU-Datenschutz-Grundverordnung und national geltender Gesetze: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2022.
 23. Dierks C, Kircher P, Engelke K, Haase M. Rechtsgutachten für das BMG - Lösungsvorschläge für ein neues Gesundheitsforschungsdatenschutzrecht in Bund und Ländern. Dierks + Company Rechtsanwalts-gesellschaft, 2019 15.09.2019. Report No.

24. NAPKON. NAPKON Nutzungsordnung [03.12.2022]. Available from: <https://cloud.napkon.de/s/kgSLW7gkBExBYfC>.
25. March S, Antoni M, Kieschke J, Kollhorst B, Maier B, Muller G, et al. [Quo Vadis Data Linkage in Germany? An Initial Inventory]. *Gesundheitswesen*. 2018;80(3):e20-e31. Epub 2018/02/21. doi: 10.1055/s-0043-125070. PubMed PMID: 29462830.
26. Bobeth C, Kleihues van-Tol K, Roessler M, Bierbaum V, Gerken M, Guenster C, et al. Methodik und Zuordnungserfolg eines Linkage von Daten klinischer Krebsregister mit Abrechnungsdaten gesetzlicher Krankenkassen. *Das Gesundheitswesen*. 2022;84:1-8. doi: 10.1055/a-1984-0085.
27. Hampf C, Geidel L, Zerbe N, Bialke M, Stahl D, Blumentritt A, et al. Assessment of scalability and performance of the record linkage tool E-PIX((R)) in managing multi-million patients in research projects at a large university hospital in Germany. *J Transl Med*. 2020;18(1):86. Epub 2020/02/19. doi: 10.1186/s12967-020-02257-4. PubMed PMID: 32066455; PubMed Central PMCID: PMCPCMC7027209.
28. Bialke M, Penndorf P, Wegner T, Bahls T, Havemann C, Piegsa J, et al. A workflow-driven approach to integrate generic software modules in a Trusted Third Party. *J Transl Med*. 2015;13:176. Epub 2015/06/05. doi: 10.1186/s12967-015-0545-6. PubMed PMID: 26040848; PubMed Central PMCID: PMCPCMC4467617.
29. Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, et al. MOSAIC--A Modular Approach to Data Management in Epidemiological Studies. *Methods Inf Med*. 2015;54(4):364-71. Epub 2015/07/22. doi: 10.3414/ME14-01-0133. PubMed PMID: 26196494.
30. Bialke M, Rau H, Thamm OC, Schuldt R, Penndorf P, Blumentritt A, et al. Toolbox for Research, or how to facilitate a central data management in small-scale research projects. *J Transl Med*. 2018;16(1):16. Epub 2018/01/27. doi: 10.1186/s12967-018-1390-1. PubMed PMID: 29370861; PubMed Central PMCID: PMCPCMC5785842.
31. Bialke M, Bahls T, Geidel L, Rau H, Blumentritt A, Pasewald S, et al. MAGIC: once upon a time in consent management-a FHIR((R)) tale. *J Transl Med*. 2018;16(1):256. Epub 2018/09/16. doi: 10.1186/s12967-018-1631-3. PubMed PMID: 30217236; PubMed Central PMCID: PMCPCMC6137912.
32. Schwaneberg T, Weitmann K, Dosch A, Seyler C, Bahls T, Geidel L, et al. Data privacy management and data quality monitoring in the German Centre for Cardiovascular Research's multicentre Translational Registry for Cardiomyopathies (DZHK-TORCH). *ESC Heart Fail*. 2017;4(4):440-7. Epub 2017/07/26. doi: 10.1002/ehf2.12168. PubMed PMID: 28742243; PubMed Central PMCID: PMCPCMC5695169.
33. Schoffer O, Rößler M, Bierbaum V, Bobeth C, Gerken M, Kleihues-van Tol K, et al. Wirksamkeit der Versorgung in onkologischen Zentren - Ergebnisbericht: Gemeinsamer Bundesausschuss - Innovationsausschuss; 2022. Available from: https://innovationsfonds.g-ba.de/downloads/beschluss-dokumente/268/2022-10-17_WiZen_Ergebnisbericht.pdf.
34. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* 2002.
35. Gesetz über die klinische und epidemiologische Krebsregistrierung im Land Nordrhein – Westfalen (Landeskrebsregistriergesetz - LKRG NRW) vom 02. Februar 2016, (2016).
36. Hampf C, Bialke M, Geidel L, Vass A, Bahls T, Blasini R, et al. A survey on the current status and future perspective of informed consent management in the MIRACUM

- consortium of the German Medical Informatics Initiative. Translational Medicine Communications. 2021;6(1). doi: 10.1186/s41231-021-00086-1.
37. March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. [Good Practice Data Linkage]. Gesundheitswesen. 2019;81(8-09):636-50. Epub 2019/08/09. doi: 10.1055/a-0962-9933. PubMed PMID: 31394579.
 38. Wikipedia. Krankenversichertennummer [6.12.2022]. Available from: <https://de.wikipedia.org/wiki/Krankenversichertennummer>.
 39. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder. Petersberger Erklärung zur datenschutzkonformen Verarbeitung von Gesundheitsdaten in der wissenschaftlichen Forschung. Entschließung der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder [Internet]. 2022; 24. November 2022. Available from: <https://www.bfdi.bund.de/SharedDocs/Downloads/DE/DSK/DSKEntschliessungen/104DSK-Petersberger-Erklaerung.pdf>.
 40. Entwurf eines Gesetzes zur Errichtung des Implantatregisters Deutschland und zu weiteren Änderungen des Fünften Buches Sozialgesetzbuch (Implantatregister-Errichtungsgesetz – EIRD), (2016).
 41. Hoffmann W, Latza U, Baumeister SE, Brunger M, Buttman-Schweiger N, Hardt J, et al. Guidelines and recommendations for ensuring Good Epidemiological Practice (GEP): a guideline developed by the German Society for Epidemiology. Eur J Epidemiol. 2019;34(3):301-17. Epub 2019/03/05. doi: 10.1007/s10654-019-00500-x. PubMed PMID: 30830562; PubMed Central PMCID: PMC6447506.
 42. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten der Deutschen Gesellschaft für Sozialmedizin und Prävention, Arbeitsgruppe Epidemiologische Methoden der Deutschen Gesellschaft für Epidemiologie, Deutsche Gesellschaft für Medizinische Informatik Biometrie und Epidemiologie, Deutsche Gesellschaft für Sozialmedizin und Prävention. [Good practice of secondary data analysis, first revision]. Gesundheitswesen. 2008;70(1):54-60. Epub 2008/02/15. doi: 10.1055/s-2007-1022529. PubMed PMID: 18273764.
 43. DGEpi. Leitlinien und Empfehlungen [12.12.2022]. Available from: <https://www.dgepi.de/de/berichte-und-publicationen/leitlinien-und-empfehlungen/>.
 44. GKV Spitzenverband. Die gesetzlichen Krankenkassen [10.12.2022]. Available from: https://www.gkv-spitzenverband.de/krankenversicherung/kv_grundprinzipien/alle_gesetzlichen_krankenkassen/alle_gesetzlichen_krankenkassen.jsp.
 45. Jungkunz M, Königeter A, Spitz M, Mehliis K, Cornelius K, Schickhardt C, et al. Stellungnahme zur Etablierung der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland. Forum Marsilius Kolleg [Internet]. 2022; 21. Available from: <https://journals.ub.uni-heidelberg.de/index.php/forum-mk/article/view/91697>.
 46. BfArM. Das FDZ Gesundheit [12.12.2022]. Available from: <https://www.forschungsdatenzentrum-gesundheit.de/das-fdz>.
 47. Reinhardt J, Hardenberg Sv, Marburger H. Kap. 5 Sozialverwaltungsverfahren | Rn. 23-26. In: Kipker D-K, Voskamp F, editors. Sozialdatenschutz in der Praxis: Nomos; 2021.
 48. Dwork C. Differential Privacy. Automata, Languages and Programming. Lecture Notes in Computer Science2006. p. 1-12.

49. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: A scoping review. *J Am Med Inform Assoc.* 2021;28(10):2269-76. Epub 2021/08/02. doi: 10.1093/jamia/ocab135. PubMed PMID: 34333623; PubMed Central PMCID: PMC8449619.
50. Boenisch F. Differential Privacy: General Survey and Analysis of Practicability in the Context of Machine Learning. Masterarbeit [Internet]. 2019. Available from: https://www.mi.fu-berlin.de/inf/groups/ag-idm/theseses/2019_Boenisch_MSc.pdf.
51. Hauer ME, Santos-Lozada AR. Differential Privacy in the 2020 Census Will Distort COVID-19 Rates. *Socius: Sociological Research for a Dynamic World.* 2021;7. doi: 10.1177/2378023121994014.
52. Santos-Lozada AR, Howard JT, Verdery AM. How differential privacy will affect our understanding of health disparities in the United States. *Proc Natl Acad Sci U S A.* 2020;117(24):13405-12. Epub 2020/05/30. doi: 10.1073/pnas.2003714117. PubMed PMID: 32467167; PubMed Central PMCID: PMC8449619.
53. Hoffmann F, Icks A. [Structural differences between health insurance funds and their impact on health services research: results from the Bertelsmann Health-Care Monitor]. *Gesundheitswesen.* 2012;74(5):291-7. Epub 2011/07/15. doi: 10.1055/s-0031-1275711. PubMed PMID: 21755492.
54. Hox J, Moerbeek M, Schoot Rvd. *Multilevel Analysis Techniques and Applications.* 3rd ed: Routledge; 2018.
55. Projektbeschreibung: VersKiK – Versorgung, Versorgungsbedarf und -bedürfnisse von Personen nach einer Krebserkrankung im Kindes- oder Jugendalter: Gemeinsamer Bundesausschuss - Innovationsausschuss; [cited 11.12.2022]. Available from: <https://innovationsfonds.g-ba.de/projekte/versorgungsforschung/verskik-versorgung-versorgungsbedarf-und-beduerfnisse-von-personen-nach-einer-krebserkrankung-im-kindes-oder-jugendalter.305>.
56. Hoffmann F, Icks A. Do persons that changed health insurance differ from those who did not? The case of diabetes. *Exp Clin Endocrinol Diabetes.* 2011;119(9):569-72. Epub 2011/08/04. doi: 10.1055/s-0031-1275277. PubMed PMID: 21811959.
57. Hoffmann F, Icks A. Diabetes prevalence based on health insurance claims: large differences between companies. *Diabet Med.* 2011;28(8):919-23. Epub 2011/03/25. doi: 10.1111/j.1464-5491.2011.03305.x. PubMed PMID: 21429008.
58. Stallmann C, Pahl R, Reuter W, Swart E. Nutzung von Leistungsdaten der Privaten Krankenversicherung in der Nationalen Kohorte am Beispiel der DKV AG. 9 Jahrestagung der Deutschen Gesellschaft für Epidemiologie; Ulm2014.
59. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten - AGENS. 14. AGENS Methodenworkshop AGENSMWS2023 2023. Available from: <https://agens.group/index.php/methodenworkshop/methodenworkshop-2023>.
60. Eijkelboom AH, de Munck L, Vrancken Peeters M, Broeders MJM, Strobbe LJA, Bos M, et al. Impact of the COVID-19 pandemic on diagnosis, stage, and initial treatment of breast cancer in the Netherlands: a population-based study. *J Hematol Oncol.* 2021;14(1):64. Epub 2021/04/19. doi: 10.1186/s13045-021-01073-7. PubMed PMID: 33865430; PubMed Central PMCID: PMC8052935.
61. Filipe MD, van Deukeren D, Kip M, Doeksen A, Pronk A, Verheijen PM, et al. Effect of the COVID-19 Pandemic on Surgical Breast Cancer Care in the Netherlands: A Multicenter Retrospective Cohort Study. *Clin Breast Cancer.* 2020;20(6):454-61. Epub 2020/09/06.

- doi: 10.1016/j.clbc.2020.08.002. PubMed PMID: 32888855; PubMed Central PMCID: PMC7413119.
62. RKI. Pseudonymisierungskonzept zur Datenübertragung im Rahmen der DaTraV (Lieferpseudonym), Version 1.8 (vertraulich). 2022.
 63. RKI. Pseudonymisierungskonzept zur Datenübertragung im Rahmen der DaTraV und für die Funktion Datenfreigabe Forschungsdatenzentrum der ePA (Periodenübergreifendes Pseudonym), Version 1.3 (vertraulich). 2022.
 64. Bundesamt für Sicherheit in der Informationstechnik. Kryptographische Verfahren: Empfehlungen und Schlüssellängen. BSI - Technische Richtlinie (BSI TR-02102-1) [Internet]. 2022.
 65. Goldhahn L, Swart E, Piedmont S. Verknüpfung von Abrechnungsdaten gesetzlicher Krankenkassen und Einsatzprotokollen des Rettungsdienstes: Brückenschlag durch Krankenversichertennummer? Das Gesundheitswesen. 2021;83(S 02):S102-S112. Epub 2021/12/02. doi: 10.1055/a-1630-7398. PubMed PMID: 34852382.
 66. DataLad [11.12.2022]. Available from: <https://www.datalad.org/>.
 67. Lee GH, Shin SY. Federated Learning on Clinical Benchmark Data: Performance Assessment. J Med Internet Res. 2020;22(10):e20891. Epub 2020/10/27. doi: 10.2196/20891. PubMed PMID: 33104011; PubMed Central PMCID: PMC7652692.
 68. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021;27(10):1735-43. Epub 2021/09/17. doi: 10.1038/s41591-021-01506-3. PubMed PMID: 34526699; PubMed Central PMCID: PMC9157510.
 69. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun. 2022;13(1):7346. Epub 2022/12/06. doi: 10.1038/s41467-022-33407-5. PubMed PMID: 36470898; PubMed Central PMCID: PMC9722782
Foley, A. Gruzdev, D. Karkada, P. Shah, J. Martin) would like to disclose the following (potential) competing interests as Intel employees. Intel may develop proprietary software that is related in reputation to the OpenFL open source project highlighted in this work. In addition, the work demonstrates feasibility of federated learning for brain tumor boundary detection models. Intel may benefit by selling products to support an increase in demand for this use-case. The remaining authors declare no competing interests.
 70. Li W, Tong J, Anjum MM, Mohammed N, Chen Y, Jiang X. Federated learning algorithms for generalized mixed-effects model (GLMM) on horizontally partitioned data from distributed sources. BMC Med Inform Decis Mak. 2022;22(1):269. Epub 2022/10/17. doi: 10.1186/s12911-022-02014-1. PubMed PMID: 36244993; PubMed Central PMCID: PMC9569919.
 71. Pfitzner B, Steckhan N, Arnrich B. Federated Learning in a Medical Context: A Systematic Literature Review. Acm Transactions on Internet Technology. 2021;21(2):1-31. doi: 10.1145/3412357. PubMed PMID: WOS:000713626000026.
 72. GKV Spitzenverband. Information des GKV-Spitzenverbandes zur Eindeutigkeit und lebenslangen Gültigkeit der Krankenversichertennummer gemäß § 290 SGB V. Information des GKV-Spitzenverbandes [Internet]. 2015. Available from: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/telematik_3/10-2015_Information_GKV-SV_Eineutigkeit_Gueltigkeit_KVNRv06.pdf.

73. DNVF. REGIBA - Fortbildung Registerbasierte Forschung [11.12.2022]. Available from: <https://www.dnvf.de/gruppen/curriculum-registerbasierte-forschung-regiba.html>.
74. PMV forschungsgruppe. Risikoevaluation und Effektivität der COVID19-Impfstoffe RiCO [14.12.2022]. Available from: <https://www.pmvforschungsgruppe.de/projekte/risikoevaluation-und-effektivitat-der-covid19-impfstoffe-rico.html>.

8 Anhang

8.1 Anlage 1a: Studienprotokoll der Proof-of-Concept Studie

Studienprotokoll

Record-Linkage von Individualdaten einer epidemiologischen Kohortenstudie (NAPKON-POP - COVIDOM Kohorte) mit GKV-Routinedatendaten: Eine Proof-of-Concept-Studie mit SARS-CoV-2 infizierten Personen nach der akuten Infektion.

Stand: 22.09.2022

Prof. Dr. Jens-Peter Reese¹, Prof. Dr. Jochen Schmitt², Dr. Peter Ihle³, Dr. Olaf Schoffer², Prof. Dr. Wolfgang Hoffmann⁴, Prof. Dr. Peter Heuschmann¹

¹Institut für Klinische Epidemiologie und Biometrie (IKEB), Julius-Maximilians-Universität Würzburg,

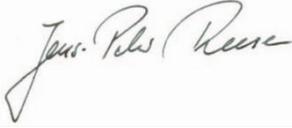
²Zentrum für Evidenzbasierte Gesundheitsversorgung (ZEGV), Universitätsklinikum Dresden Carl Gustav Carus

³PMV Forschungsgruppe, Universität zu Köln

⁴Institut für Community Medicine / Abt. Versorgungsepidemiologie und Community Health, Universität Greifswald

VERTRAULICH

Die Informationen in diesem Studienprotokoll sind streng vertraulich zu behandeln. Sie dienen nur zur Information von Studienmitarbeitern:innen, der Ethikkommission und zuständigen Behörden. Dieses Studienprotokoll darf ohne Zustimmung der Studienleiter nicht an Unbeteiligte weitergegeben werden.

Titel	Record-Linkage von Individualdaten einer epidemiologischen Kohortenstudie (NAPKON-POP-COVIDOM Kohorte) mit GKV-Routinedatendaten: Eine Proof-of-Concept-Studie mit SARS-CoV-2 infizierten Personen nach der akuten Infektion.
Studienleiter	<p>Prof. Dr. Jens-Peter Reese</p>  <hr/> <p>Unterschrift Prof. J.-P- Reese</p> <p>Anschrift: Julius-Maximilians-Universität Würzburg Institut für Klinische Epidemiologie und Biometrie Petrinistr. 33a 97080 Würzburg</p> <p>E-Mail: Reese_j@ukw.de</p> <p>Telefonnr.: 0931 – 201 47322</p>
Beteiligte und Aufgaben	<p>Prof. Dr. Jochen Schmitt, Dr. Olaf Schoffer, Zentrum für evidenzbasierte Gesundheitsversorgung (ZEG)V, Universitätsklinikum Dresden</p> <ul style="list-style-type: none"> • Aufgaben der Datenannahme- und Aufbereitungsstelle am ZEGV Dresden: • Annahme und Zusammenführen der verschiedenen KK-Daten • Harmonisierung der verschiedenen Krankenkassen (KK)-Daten hinsichtlich Datenformat und Variablendefinitionen • Plausibilitäts- und Konsistenzprüfung • Selektion und Pseudonymersetzung der für das Forschungsvorhaben benötigten Daten • Übermittlung des pseudonymisierten Datenauszugs an die Auswertestelle <p>Erfahrungen des ZEGV Dresden:</p> <p>Das ZEGV ist eine der deutschlandweit führenden Stellen für Methodik und Umsetzung von Versorgungsforschung auf verschiedensten versorgungsnahen Daten. Insbesondere werden hier seit etlichen Jahren gesetzliche Krankenversicherung (GKV)-Abrechnungsdaten verschiedener Krankenkassen aufbereitet und analysiert [4-6], nicht zuletzt für Post-COVID-Fragestellungen [9]. Im Rahmen des vom ZEGV geleiteten Wirksamkeit der Versorgung in onkologischen Zentren (WiZen)-Projektes wurden zudem Daten verschiedener Krebsregister harmonisiert und erfolgreich auf Individualdatenebene mit bundesweiten GKV-Daten verknüpft.</p>

	<p>Prof. Dr. Jens-Peter Reese, Peter Heuschmann, IKEB, Universität Würzburg</p> <p>Analysestelle für verknüpfte Daten:</p> <ul style="list-style-type: none"> • Verknüpfung der übermittelten Datenquellen • Auswertung der verknüpften Datenquellen gemäß der im Studienprotokoll genannten Fragestellungen <p>Dr. Peter Ihle, PMV, Universität zu Köln,</p> <p>Beratung zur Annotation, Kodierung und Kategorisierung, Erstellung von Metadatenverzeichnissen, Beratung zur Validität der übermittelten Items, angemessenen Analysemethodik, Übertragbarkeit der Analyseergebnisse</p> <p>Prof. Wolfgang Hoffmann, Universität Greifswald</p> <p>Beratung für Datenflüsse, Konzepte zu Datenübermittlung, Erhebung und Management modularer Patienteneinwilligungen, Datenschutz und Datensicherheit; Datenanalyse</p>
Studiendesign	<p>Proof-of-Concept-Studie der Versorgungsforschung – Retrospektive Verknüpfung von klinischen Primärdaten einer prospektiven Kohortenstudie und Individual- GKV-Routinedaten.</p>
Hintergrund	<p>Seit Dezember 2019 breitet sich das neue Coronavirus SARS-CoV-2 weltweit aus. Dieses Ausbruchsgeschehen wird seit dem 11.03.2020 von der Weltgesundheitsorganisation (WHO) als Pandemie klassifiziert. Um den akuten Verlauf der Covid 19-Krankheit zu erfassen, aber auch um deren langfristigen Folgen zu analysieren, wurden im Rahmen der im Rahmen der Bundesministerium für Bildung und Forschung (BMBF)-geförderten Initiative Netzwerk Universitätsmedizin (NUM, https://www.netzwerk-universitaetsmedizin.de/) durchgeführten Nationalen Pandemie Kohorten Netzwerk (NAPKON)-Projekts (für Details siehe Anhang 1 und [1]) drei Kohortenstudien etabliert, die verschiedene Abschnitte im Covid-19 Krankheitsverlaufs abbilden können: NAPKON-POP (populationsbasierte Plattform), -SÜP (sektorenübergreifende Plattform) und -HAP (hochauflösende Plattform). Diese individuell erhobenen klinischen Daten liefern präzise Informationen zu Diagnostik, Laborbefunden, Schweregradeinteilungen von COVID-19 und Symptomen sowie patient-reported outcome measures (PROMs).</p> <p>Eine informed-consentbasierte, personenbezogene Verlinkung dieser wichtigen klinischen Individualdaten mit GKV-Routinedaten würde eine weitaus umfassendere Analyse von Krankheits- und Versorgungsprozessen bei COVID-19 Infektionen ermöglichen, als dies alleine mit klinischen Daten möglich wäre. GKV-Routinedaten ermöglichen die vollständige und sektorenübergreifende Abbildung von Krankheitsverläufen inkl. Angaben zu Vorerkrankungen sowie langfristigen Nachbeobachtungszeiträumen (bis zu mehreren Jahren). GKV-Routinedaten bilden eine umfassende, effizient nutzbare,</p>

	<p>qualitativ hochwertige Grundlage für die Untersuchung der Inanspruchnahme, Risikofaktoren, Erkrankungslast, Pflegebedürftigkeit, Versorgungswege, und der Wirksamkeit, Sicherheit und Kosten von medikamentösen Therapien und Versorgungskonzepten [1]. Auf der Basis hochwertiger deskriptiver Evidenz aus der Versorgungsforschung können in der Folge wirksame Versorgungskonzepte für die COVID-19 Erkrankung entwickelt und implementiert werden.</p>
<p>Ziele und Fragestellungen</p>	<p>Ziel dieser Studie ist es, die Machbarkeit der consentbasierten Datenverknüpfung als Proof-of-Concept (PoC) zu bestätigen, harmonisierte Prozesse des gesamten Datenflusses zu etablieren und eine erste ergänzende orientierende Validierung der verknüpften Daten durchzuführen. Die gewonnenen Erkenntnisse sollen für den Rollout über alle NAPKON-Kohorten genutzt werden.</p> <p>Die Verknüpfung von GKV-Routinedaten mit NAPKON Daten bietet die einzigartige Möglichkeit, ein umfassendes Bild der Gesundheitsversorgung der NAPKON-StudienTeilnehmende vor und nach der eigenen Sars-Cov2-Infektion sowie vor und nach der Pandemie und zu zeichnen. Insbesondere (Ko-)Morbidity, Medikamente, Verordnungen von Heil- und Hilfsmitteln, stationäre Aufenthalte und ambulante Behandlungen können über einen längeren Zeitraum abgebildet werden. Anhand der klinischen Daten wiederum können erweiterte Risikoadjustierungen zu einem vollständigeren Bild führen.</p> <p>Um eine übergreifende Nutzung von GKV-Routinedaten durch individuelles Datenlinkage in den drei NAPKON Kohorten nutzbar zu machen wird zunächst in einem PoC-Ansatz für NAPKON-POP entwickelt und in weiteren Schritten für alle NAPKON Kohorten ausgerollt:</p> <p>Ziele Proof-of-Concept Phase</p> <p>Im Rahmen dieser Studie sollen folgende Fragestellungen untersucht werden:</p> <ol style="list-style-type: none"> 1. Ist eine Datenverknüpfung von GKV-Routinedaten kooperierender Krankenkassen mit den NAPKON-Pop-Kohortendaten möglich? 2. Welcher Anteil an NAPKON-Pop StudienTeilnehmende willigt in die Nutzung und Verknüpfung von GKV-Daten ein? Unterscheiden sich die Einwilliger systematisch von Nicht-Einwilligern? Falls ja, für welche Variablen bestehen wesentliche Unterschiede? Sind ggf. Selektionseffekte durch Nicht-Einwilligung definierbarer Personengruppen zu erwarten? 3. Welche Ergänzungen der bestehenden NUM-Infrastruktur sind notwendig, um die Voraussetzungen für ein datenschutzkonformes und praktikabel einsetzbares

	<p>Datenlinkage von GKV-Daten und primären Kohortendaten zu schaffen?</p> <ol style="list-style-type: none"> 4. Welche prozessbezogenen Hindernisse bestehen bzgl. des vorgesehenen Datenlinkage? 5. Wie groß ist der personelle, zeitliche und IT-technische Aufwand für die (routinemäßige) Umsetzung? 6. Zu welchem Grad gelingt es, mit den GKV-Daten relevante Lücken in dem vor allem allein durch die epidemiologischen Daten gezeichneten Bild zu füllen? 7. In welchen inhaltlichen Bereichen und zu welchem Grad können die erhobenen Daten zum Krankheitsverlauf und zum Versorgungsgeschehen beider Datenquellen wechselseitig validiert werden? 8. Können bestehende Definitionen von "Long-Covid", die mit klinischen Individualdaten dargestellt werden, ebenso mittels GKV-Routinedaten abgebildet werden? 9. Welche Variablen und welche operationalen Schritte sind bestehende Definitionen von "Long-Covid" nötig und welche davon sind möglich? <p>Ziele Rollout Phase:</p> <p>Nach erfolgreicher Umsetzung und Auswertung der PoC-Phase in der NAPKON-POP Kohorte werden die entwickelten Prozesse auf alle drei NAPKON-Kohorten ausgerollt, um Forschenden einen verknüpften GKV-Routinedatensätze für einen möglichst großen Anteil der klinischen Individualdaten zur Verfügung stellen zu können. Die Rollout-Phase wird begleitet durch ein engmaschiges Berichtssystem innerhalb und Erfahrungsaustausch zwischen den NAPKON-Kohorten. Auf diese Weise sollen möglichst frühzeitig notwendige Schnittstellenprobleme erkannt werden und Vorschläge für ein harmonisiertes Vorgehen entwickelt werden.</p>
Fallzahl und Analyse	<p>Die Fallzahlbegründung basiert auf Fragestellung (1) der Machbarkeitsstudie: Ist eine Datenverknüpfung von GKV-Routinedaten kooperierender Krankenkassen mit den NAPKON-Pop-Kohortendaten möglich?</p> <p>Für diese engere Fragestellung kann eine geringe Fallzahl bereits ausreichend sein. In Anlehnung an Viechtbauer et. al. (2015) - Berechnung des Stichprobenumfangs in Pilotstudien - wird die Fallzahl wie folgt begründet:</p> <p>Zunächst wird angenommen, dass ein (noch undefiniertes) Problem bei der Datenverknüpfung mit einer Wahrscheinlichkeit von 15% auftritt (z. B. nicht korrekte Schlüsselvariablen o.Ä.). Um mit einem Vertrauensniveau von 95% ein solches Problem zu erkennen und unter der Annahme, dass bei 15% der Verknüpfungen das Problem auftritt, werden mindestens 18.4 Proband:innen benötigt.</p>

	$n = \frac{\ln(1-\gamma)}{\ln(1-\pi)} = \frac{\ln(1-0.95)}{\ln(1-0.15)} = 18,43$ <p>γ = Vertrauensniveau 95%; π = Fehlerwahrscheinlichkeit 15%</p> <p>Bei 19 Proband:innen kann man mit einer 95% Sicherheit davon ausgehen, dass zumindest einmal ein Verknüpfungsproblem identifiziert wird. Für die POP Kohorte Standort Würzburg wird diese Anzahl von n= 19 von 5 Krankenkassen überschritten, die für eine Kooperation kontaktiert und angefragt wurden.</p> <p>Für den Standort Berlin (zum Teil andere Kassen) ist eine ähnliche Anzahl zu erwarten. Die Zahlen in Kiel können höher liegen.</p> <p>Die Studie gewinnt an Aussagekraft, wenn Proband:innendaten aller Studienzentren (Würzburg, Berlin, Kiel) verlinkt werden bzw. wenn möglichst viele unterschiedliche GKVen eingeschlossen werden können.</p> <p>Die Auswahl der GKVen basiert zunächst auf der Anzahl der in NAPKON-POP eingeschlossenen Proband:innen dieser GKVen.</p> <p>Die Beantwortung quantitativer Fragestellungen (2 in Teilen 6,7,8,9) der Machbarkeitsstudie gewinnt an Aussagekraft bzw. Präzision, wenn möglichst viele der NAPKON-POP Datensätze in die Verknüpfung und ein anschließende Analyse einfließen können. Zur Überprüfung der Machbarkeit liegen derzeit ca. von 600 Proband:innen unterschiedlicher Kassen an allen drei Standorten vor.</p> <p>Quantitative Analysen der Machbarkeitsstudie sind zunächst deskriptiv. Es werden die zugehörigen Metriken (Anteile, Zentralmaße und Verteilungsmaße) berechnet und berichtet.</p>
Datengrundlage	<p>Datengrundlage für die Datenverknüpfung und nachfolgende Analyse bilden</p> <p>(1) Identifikation (ID)-Datender Kohorten des NAPKON in der Unabhängigen Treuhandstelle (uTHS) des NUKLEUS (NUM Klinische Epidemiologie und Studienplattform) des NUM (Infrastrukturen siehe unten Anhang) und</p> <p>(2) ID-Daten (z. B. die GKV-Nr.) der ambulanten und stationären Verordnungs- und Abrechnungsdaten ausgewählter GKVen. Die ID-Daten werden ausschließlich der uTHS übergeben und nach Zusammenführung mit den Kohorten-Teilnehmende pseudonymisiert. In die Datenhaltung werden lediglich die pseudonymisierten Datensätze übertragen. Für die Auswertung stehen ausschließlich die in der uTHS pseudonymisierten Daten zur Verfügung.</p> <p>zu (1) Proof-of-Concept Phase: NAPKON-POP Kohorte (Details siehe Anlage 1 -3)</p>

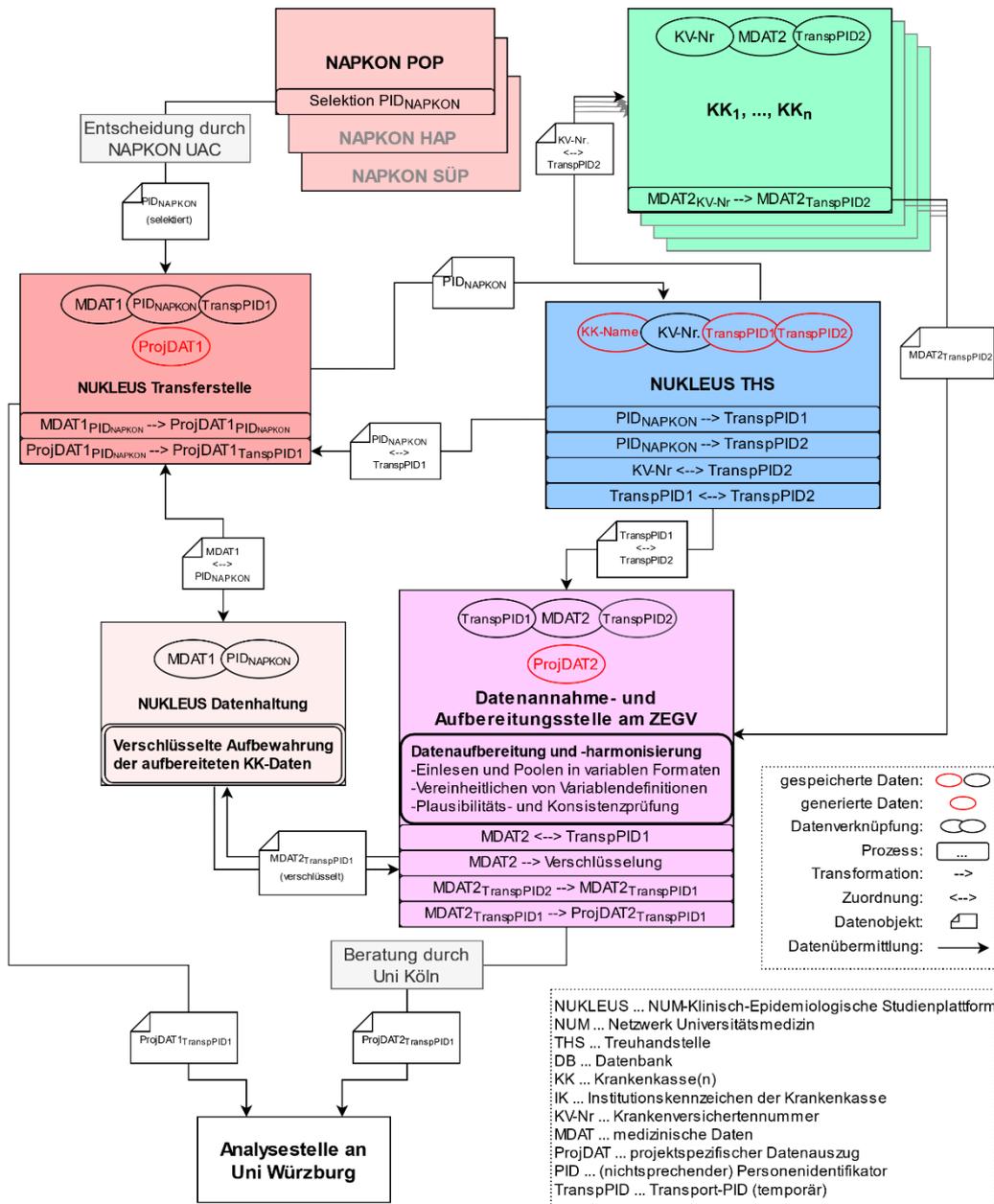
	<p>In der NAPKON-POP [2] werden in geographisch definierten Sammelgebieten alle Patientinnen und Patienten mit überstandener SARS-CoV-2-Infektion schweregradunabhängig und populationsrepräsentativ identifiziert und in ein diagnostisches Langzeitprogramm eingeladen. Es erfolgt die detaillierte retrospektive Erfassung des akuten Erkrankungsverlaufes und eine prospektive Biomaterialsammlung sowie die Erfassung der Langzeitfolgen und Komorbiditäten</p> <p>Insgesamt sind während der ersten Förderperiode (BMBF FKZ 01KX2021) von NAPKON-POP im Rahmen von NAPKON bis 31.12.2021 über 2000 Patientinnen und Patienten an den drei Standorten rekrutiert werden (Kiel: >1663, Berlin: 302, Würzburg 348). In der zweiten Förderperiode (2022-2024) sollen weitere ca. 1.200 Patientinnen und Patienten an den drei Standorten rekrutiert werden (Kiel: ca. 700, Berlin: ca. 250, Würzburg ca. 250).</p> <p><i>Einschlusskriterien</i></p> <ul style="list-style-type: none"> • Nachgewiesene SARS-CoV-2 Infektion mindestens 6 Monate zurückliegend (Molekularer Nachweis von SARS-CoV-2 aus Abstrich oder Körpersekret) • Erstwohnsitz in einer der definierten Studienregionen • Alter ≥18 Jahre • schriftliche oder elektronische Einwilligung des Proband:innen liegt vor <p><i>Ausschlusskriterien</i></p> <ul style="list-style-type: none"> • Akute Re-Infektion mit SARS-CoV-2 <p>ANHANG 2 zeigt die Elemente der NAPKON-POP-Erhebung und -Untersuchung. Nach der Basisuntersuchung sind 12 monatliche Nachuntersuchungen für ausgewählte Patientinnen und Patienten mit anhaltenden COVID-19-Beschwerden vorgesehen.</p> <p>Zu 2: Die zweite Datengrundlage bilden pseudonymisierte ambulante und stationäre Verordnungs- und Abrechnungsdaten ausgewählter GKVen.</p> <p>Folgende spezifische Daten sind zur Beantwortung der Forschungsfragen notwendig:</p> <ul style="list-style-type: none"> • Stammdaten der Versicherten: Pseudonummer, Krankenkassennummer, Geschlecht, Geburtsjahr, Ende der Versichertenzeit, Postleitzahl des Versicherten, Abgangsgrund, Todestag • Versichertenzeiten des Versicherten, Beginn, Ende, Versichertenart (kategorisiert), Tätigkeitsschlüssel • Versichertenmerkmale: Merkmalschlüssel, gültig ab, gültig bis • Daten des Behandlungsscheins: Pseudonummer des Versicherten, Abrechnungsjahr, Abrechnungsquartal, Abrechnungsjahr und -quartal, Kassenzuständige Krankenversicherung (KV)-ID, KV-ID des abrechnenden Arztes, Fallnummer des Behandlungsscheines, Pseudonym der Betriebsstätte des Behandlungsscheines, Pseudonym der überweisenden Betriebsstätte, Kennzeichnung der Art der Inanspruchnahme, Unfallkennzeichen, Behandlungsart,
--	---

	<p>Anzahl Punktzahlzehntel, Inhalt des Feldes KOSTEN, Dialysesachkosten in Cent, Behandlungsbeginn, Behandlungsende, Pseudonym der lebenslangen Arztnummer des überweisenden Arztes, Arztgruppenschlüssel der behandelnden Ärzte, Arztgruppenschlüssel des überweisenden Arztes, Pseudonym der lebenslangen Arztnummer des behandelnden Arztes, Kennzeichen Bereichseigen oder -fremd</p> <ul style="list-style-type: none"> • Ambulant abgerechnete Gebührenordnungspositionen: Behandlungsjahr, Behandlungsquartal, Fallnummer, Kassenzuständige KV-ID, KV_ID des abrechnenden Arztes, Pseudonym der Betriebsstätte, Gebührenordnungsnummer, Datum der Leistungserbringung, Kennzeichen Bereichseigen oder -fremd • Ambulante Diagnosen: Behandlungsjahr, Behandlungsquartal, KV_ID des abrechnenden Arztes, Fallnummer, Diagnosesicherheit, Lokalisation, ICD Schlüssel der Diagnose, • Pseudonym der Betriebsstätte des Arztes, Bereichseigen / fremd • Ambulante Operationen: Behandlungsjahr, Behandlungsquartal, Kassenzuständige KV-ID, KV_ID des abrechnenden Arztes, Fallnummer, Lfd. Nummer je Fall, OPS-Schlüssel, Lokalisation, Pseudonym der Betriebsstätte des Arztes • Ambulant verordnete Arzneimittel – Kopfdaten: Pseudonummer des Versicherten, Identifikationsnummer der Verordnung, Zeitraum der Arzneimittelabrechnung, Pseudonym Lebenslange Arztnummer (LANR), Arztgruppenschlüssel der LANR (ab Q3 2008), Ausstellungsdatum der Verordnung durch den Arzt, Abgabedatum durch die Apotheke, Kennzeichen Zuzahlungsbefreiung, Kennzeichnung, ob ein BVG-Fall vorliegt, Kennzeichnung, ob die Verordnung als Sprechstundenbedarf abgerechnet wurde, Kassenummer, Jahr- und Quartalsangabe der Verordnung • Ambulant verordnete Arzneimittel – Positionen: Identifikations-nummer der Verordnung, Zeitraum der Arzneimittelabrechnung, Pharmazentralnummer 7-stellig oder Hilfsmittelnummer 10-stellig, Faktor, Bruttobetrag, Preis Rezept, Jahr und Quartal • Stationäre Abrechnungsdaten – Fallinformationen: pseudonymisierte Nummer des Krankenhausfalles, Pseudonummer des Versicherten • Krankenhaus: pseudonymisierte IK-Nummer, Bettenzahl des Hauses, Krankenhaus Universitätsklinikum ja oder nein, Aufnahmedatum, Entlassungsdatum, Aufnahmegrund, Aufnahmearbeit, Entlassungsgrund, Entlassungsanlass, Diagnosis Related Group (DRG) oder pauschalierendes Entgeltsystem für Psychiatrie und Psychosomatik (PEPP), Belegungstage, Bruttoausgaben, Hauptdiagnose • Stationäre Abrechnungsdaten – Operationsschlüssel: pseudonymisierte Nummer des Krankenhausfalles,
--	--

	<p>Operationen- und Prozedurenschlüssel (OPS)/Prozeduren, Lokalisation einer Prozedur</p> <ul style="list-style-type: none"> • Stationäre Abrechnungsdaten – Diagnosen: pseudonymisierte Nummer des Krankenhausfalles, Diagnosestellungsart (Haupt- oder Nebendiagnose), Primärdiagnose, Sekundärdiagnose, Lokalisation einer Primärdiagnose, Lokalisation einer Sekundärdiagnose • Stationär erbrachte ambulante Abrechnungsdaten – Fallinformationen: pseudonymisierte Nummer des Krankenhausfalles, Pseudonummer des Versicherten • Stationäre erbrachte ambulante Abrechnungsdaten – Operationsschlüssel: pseudonymisierte Nummer des Krankenhausfalles, OPS/Prozeduren, Lokalisation einer Prozedur • Stationär erbrachte ambulante Abrechnungsdaten - Diagnosen: pseudonymisierte Nummer des Krankenhausfalles, Diagnosestellungsart, Primärdiagnose, Sekundärdiagnose, Lokalisation einer Primärdiagnose, Lokalisation einer Sekundärdiagnose • Arbeitsunfähigkeitsfall: Fallnummer, Pseudonummer des Versicherten, Arbeitsunfähigkeit ab, Arbeitsunfähigkeit bis • Diagnosen zum Arbeitsunfähigkeitsfall: Fallnummer, Anfangsdatum Diagnosegültigkeit, Enddatum Diagnosegültigkeit, International Classification of Diseases (ICD) Schlüssel • Leistungsabrechnung Heil- und Hilfsmittel: Pseudonummer des Versicherten, Verordnungs-/Ausstellungsdatum, Pseudonym Arztnummer, Heilmittelindikation, Menge, Anzahl Gebührenpositionen, Gebührenposition, Bruttobetrag, Leistung von, Leistung bis, Leistungsbereich, Verordnungsart Heilmittel • Teilnahme der Versicherten an Kassenverträgen: Pseudonummer des Versicherten, Teilnahmebeginn des Versicherten, Teilnahmeende des Versicherten, Kennzeichnung des Vertrages • Pflegegrad/Pflegestufe: Pseudonummer des Versicherten, Pflegegrad/Pflegestufe, Gültig ab, Gültig bis <p>Mehrfach genannte Variablen dienen zur Verknüpfung der verschiedenen Datentabellen.</p>
Datenmanagement und Datenschutz	<p>Die ID-Daten und pseudonymisierten medizinischen Daten der NAPKON-Kohorte werden gemäß den Studienprotokollen erhoben und verarbeitet, die durch zuständige Ethikkommissionen begutachtet wurden. Die Verwendung und Zusammenführung der ID-Daten in der uTHSist durch die Patienteninformationen und Einverständniserklärungen der NAPKON-Kohorten abgedeckt. Die verwendeten modularen Einwilligungserklärungen finden sich in ANHANG 5.</p> <p>Für die NAPKON-POP Standorte liegen derzeit modulare Einwilligungen zur Datenzusammenführung, Kontakt- und Versicherungsdaten für ca. 600 Proband:innen sowie ein positives</p>

	<p>Ethikvotum der Medizinischen Fakultät der Universität Würzburg vor (236/20_z).</p> <p>Die Datennutzung erfolgt gemäß der NAPKON Nutzungsordnung und den darin festgelegten Use & Access Regularien. Entsprechend wird vor der Datenverknüpfung mit GKV-Daten die Datennutzung beim NAPKON Use und Access Komitee (UAC) beantragt. Die Datenherausgabe der NAPKON erfolgt über eine Datentransferstelle (für Details siehe Diagramm (Abbildung 2)).</p> <p>Das Datenlinkage erfolgt dabei unter Einbeziehung einer Treuhandstelle (THS), welche den Identitätsabgleich zwischen den verschiedenen Datenquellen vornimmt, jedoch im Sinne des Trennungsgebots keine darüber hinaus gehenden Daten erhält und die identifizierenden Angaben auch nicht weitergibt. An die Datenannahme- und Aufbereitungsstelle wird ausschließlich eine Zuordnung von Transport-Pseudonymen weitergegeben.</p> <p>Zur Gewährleistung des Datenschutzes wurden ZEGV, TU Dresden getrennte Stellen für Datenmanagement und Datenauswertung ohne gegenseitige Weisungsberechtigung etabliert. Die Arbeitsgruppe Datenmanagement nimmt die pseudonymisierten Daten der Krankenkassen und die Pseudonyme aus der uTHS in Empfang, poolt, harmonisiert und prüft diese hinsichtlich Plausibilität bzw. Konsistenz, bereitet sie gemäß der spezifischen Fragestellung des Projekts auf und gibt ausschließlich anonymisierte Daten als Auszug aus den bereitgestellten Versorgungsdaten der Krankenkassen an die Auswertestelle weiter.</p> <p>Die beteiligten Projektmitarbeitende des ZEGV arbeiten auf der Grundlage eines einrichtungsspezifischen Datenschutzkonzepts, das auf die Einhaltung der datenschutzrechtlichen Bestimmungen, sowie der guten wissenschaftlichen Praxis und der guten Praxis Sekundärdatenanalyse [7] abzielt. Die Einhaltung der Standard Operating Procedure (SOP)s Datenschutz und Datensicherheit sind für alle Projektmitarbeitende obligatorisch. Die SOP enthält u. a. Angaben zur personellen Verantwortlichkeit, zum Zugriffsschutz, zur Datensicherung und zur Verfügbarkeitskontrolle. Die Sicherung und der Schutz von Sozialdaten sind damit nach den Anforderungen gemäß § 78a SGB X geregelt.</p> <p>Zudem werden die Daten nur anlassbezogen verarbeitet und es erfolgt keine permanente Datenhaltung der unverschlüsselten Daten außerhalb der jeweiligen Datenlieferanten (NAPKON/NUM NUKLEUS bzw. KK).</p>
--	---

Abbildung 1 : Datenfluss, Datenverknüpfung und Datenanalyse



<p>Datenfluss, Datenverknüpfung und Datenanalyse</p>	<p>In Abbildung 2 sind Datenfluss, Datenverknüpfung und Datenanalyse schematisch abgebildet.</p> <p>Üblicherweise können die notwendigen Sekundärdaten von den angefragten Krankenkassen mit einer Latenz von 1 bis 4 Monaten für die stationären und von 7 bis 9 Monaten für die ambulanten Daten bereitgestellt werden. Bei einem vorgesehenen Linkage verfügbarer Daten im 4. Quartal 2022 können somit umfassende Analysen für</p>
--	--

<p>einen Zeitraum vor Pandemiebeginn (ab Quartal 1, 2019) für das gesamte Jahr 2021 durchgeführt werden.</p> <p>Einschluss von Patient*innen in NAPKONPOP mit Einwilligung zur Erfassung der Krankenkassendaten:</p> <p>Sobald ein:e Patient:in in der NAPKONPOP Kohorte in die Verknüpfung mit den Krankenkassendaten einwilligt, wird bei der NUKLEUS THS zusätzlich die KV-Nr. des:der Patient:in und ein Kennzeichen der jeweils versichernden KK erfasst (z. B. Name der Kasse).</p> <p>Anfrage und Herausgabe der Krankenkassendaten einer spezifizierten Kohorte von Patientinnen und Patienten:</p> <p>Die NAPKON POP oder andere Forscher:innen stellen über das NAPKON UAC einen Antrag zur Herausgabe von Krankenkassendaten zu einer spezifizierten Kohorte/selektierte Teilnehmende.</p> <p>Nach positivem Votum des UAC wird über die NUKLEUS Transferstelle (Sabine Hanß; Institut für Medizinische Informatik der Universitätsmedizin Göttingen) für selektierte Befragungsteilnehmer/Patientinnen und Patienten bei der THS nach dem Einwilligungstatus zur Datenverknüpfung über das das zugehörige NAPKON-Pseudonym (PID_{NAPKON}) bei der Treuhandstelle erfragt.</p> <p>Bei aktuell vorliegender Einwilligungserklärung in die Datenverknüpfung erzeugt die THS aus dem NAPKON-Pseudonym zwei unterschiedliche systemfreie personenbezogene Transport-IDs (TranspPID1, TranspPID2), ordnet die in der Treuhandstelle vorliegende KV-Nr. der TranspPID2 zu und übermittelt die KV-Nr. und die TranspPID2 inklusive Zuordnungsvorschrift an die jeweils versichernde KK. Zudem übermittelt die THS die TranspPID1 und die TranspPID2 inklusive Zuordnungsvorschrift an die Datenannahme- und Aufbereitungsstelle am ZEGV Dresden sowie die PID_{NAPKON} und TranspPID1 inklusive Zuordnungsvorschrift an die NUKLEUS Datenhaltung. Die THS übernimmt die Ansprache der KKen und steht für Rückfragen von den KKen zur Verfügung Die Übermittlungswege zu den KKen, sowie die zu übermittelnden Daten an die KKen werden kassenspezifisch mit jeder Einrichtung auf Basis ihrer datenschutzrechtlichen Vorgaben vereinbart.</p> <p>Die Krankenkassen selektieren die Daten (MDAT2) der bei ihnen versicherten BefragungsTeilnehmende mit Einwilligung zur Datenverknüpfung anhand der KV-Nr., ersetzen die KV-Nr. durch die TranspPID2 und übermitteln diesen pseudonymisierten Datensatz an die Datenannahme- und Aufbereitungsstelle.</p> <p>Die Datenannahme- und Aufbereitungsstelle (ZEGV, Hochschulmedizin Dresden) führt die Daten der verschiedenen KKen zusammen und ersetzt anhand der vorliegenden Zuordnungsvorschrift die TranspPID2 durch die TranspPID1. Zudem werden die aus den</p>
--

	<p>verschiedenen KK mit unterschiedlichem Datenformat bereitgestellten Daten hinsichtlich Datenformat und Variablendefinitionen harmonisiert, auf Plausibilität und Konsistenz überprüft und die für das Forschungsvorhaben benötigten Variablen zur Weitergabe selektiert (ProjDAT2). Um keine dauerhafte Datenhaltung der MDAT2 am ZEGV zu etablieren, aber auf Rückfragen und Nachforderungen ohne erneute Datenlieferung der unterschiedlichen KK angemessen reagieren zu können, werden nach Extraktion der ProjDAT2 die aufbereiteten und harmonisierten MDAT2 mit zugeordneter TranspPID1 verschlüsselt an die NUKLEUS Datenhaltung übergeben und im ZEGV anschließend gelöscht. Der Schlüssel zum Entschlüsseln der Daten verbleibt im ZEGV. So hat keine beteiligte Stelle dauerhaften Zugriff auf die aufbereiteten MDAT2, können aber anlassbezogen die verschlüsselten Daten an das ZEGV rückübermittelt, dort entschlüsselt und für weitere Datenaufbereitungen genutzt werden</p> <p>Die PoC-Analyse wird anhand der verknüpften Daten nach Entscheidung des NAPKON UAC bezüglich der Nutzung von NAPKON-POP-Daten und nach Beratung durch die Universität Köln bezüglich der Nutzung von KK-Daten an der Analysestelle an der Universität Würzburg durchgeführt. Die NAPKON NUKLEUS Datenhaltung sowie die Datenannahme- und Aufbereitungsstelle erstellen jeweils einen auf das Forschungsprojekt zugeschnittenen Auszug der jeweils vorliegenden NAPKON-POP- bzw. KK-Daten (MDAT1 und MDAT2, jeweils pseudonymisiert mit Bezug zu TranspPID1) und übermitteln diese an die Analysestelle. Diese kann anhand der gleichartigen TranspPID1 die beiden Datenquellen verknüpfen und gemäß der Use&Access-Entscheidung und Studienprotokoll verknüpft auswerten.</p> <p>Alle beschriebenen Datenflüsse zwischen verschiedenen beteiligten Stellen erfolgen geschützt (verschlüsselt und/oder über speziell geschützte Transportwege).</p> <p>Aufgaben der beteiligten Stellen:</p> <p>NAPKON-POP (Thomas Bahmer, Stefan Schreiber) /Forschungsprojekt</p> <ul style="list-style-type: none"> • Selektion der Teilnehmende mit Einwilligung zur Datenverknüpfung <p>NAPKON-POP / NUKLEUS Datenhaltung und Transferstelle (Sabine Hanß; Institut für Medizinische Informatik der Universitätsmedizin Göttingen):</p> <ul style="list-style-type: none"> • Übermittlung des NAPKON-Pseudonyms an die THS • Erstellung eines projektspezifischen Datenauszugs • Ersetzung des NAPKON-Pseudonyms durch die TranspPID1 für diesen Datenauszug • Übermittlung des neu pseudonymisierten Datenauszugs an die Auswertestelle
--	---

	<ul style="list-style-type: none"> • Aufbewahrung der aufbereiteten und verschlüsselten KK-Daten • Anlassbezogene Rückübermittlung der aufbereiteten und verschlüsselten KK-Daten an ZEGV <p>NUKLEUS Treuhandstelle:</p> <ul style="list-style-type: none"> • Ansprache der KKen • Erzeugung der TranspPID1 und TranspPID2 aus dem NAPKON-Pseudonym • Zuordnung der KV-Nr. zur TranspPID2 und Übermittlung beider inklusive Zuordnungsvorschrift an die jeweils versichernde KK (gemäß KK-Kennzeichen und einrichtungsbezogener Vorgaben zur Übermittlung) • Übermittlung von TranspPID1 und TranspPID2 inklusive Zuordnungsvorschrift an die Datenannahme- und Aufbereitungsstelle • Übermittlung von PID_{NAPKON} und TranspPID1 inklusive Zuordnungsvorschrift an die NUKLEUS Datenhaltung <p>Krankenkassen:</p> <ul style="list-style-type: none"> • Erstellung eines Datenauszugs anhand der übermittelten KV-Nr. und gemäß der Variablenliste aus dem §75-Antrag • Ersetzen der Krankenversicherungsnummer durch die TranspPID2 im Datenauszug • Übermittlung dieses pseudonymisierten Datenauszugs an die Datenannahme- und Aufbereitungsstelle <p>Datenannahme- und Aufbereitungsstelle am ZEGV Dresden:</p> <ul style="list-style-type: none"> • Annahme und Zusammenführen der verschiedenen KK-Daten • Ersetzen der TranspPID2 durch die TranspPID1 • Harmonisierung der verschiedenen KK-Daten hinsichtlich Datenformat und Variablendefinitionen • Prüfung auf Plausibilität und Konsistenz • Selektion der für das Forschungsvorhaben benötigten Variablen • Übermittlung des pseudonymisierten Datenauszugs an die Auswertestelle • Verschlüsselung, Übermittlung an NUKLEUS Datenhaltung und Löschung der aufbereiteten KK-Daten • Anlassbezogene Entschlüsselung der rückübermittelten aufbereiteten und verschlüsselten KK-Daten für weitere Datenaufbereitungen <p>NAPKON Use & Access-Board</p> <ul style="list-style-type: none"> • Entscheidung über Nutzung der NAPKON POP-Daten für PoC-Forschungsvorhaben <p>Universität Köln (PMV)</p> <ul style="list-style-type: none"> • Beratung der Auswertestelle hinsichtlich der Nutzung von KK-Daten
--	--

	<p>Auswertestelle an der Universität Würzburg (IKEB)</p> <ul style="list-style-type: none"> • Verknüpfung der übermittelten Datenquellen anhand der TranspPID1 • Auswertung der verknüpften Datenquellen gemäß Studienprotokoll <p>Die vorbereiteten klinischen Daten aus der NAPKON-POP Kohorte werden entsprechend der Use& Access Regularien über die NAPKON Datentransferstelle pseudonymisiert ebenfalls an die Datenanalysestelle übermittelt. Hier erfolgt die Datenverknüpfung.</p> <p>Die nachfolgende Datenanalyse (in der Roll-Outphase oder nachfolgenden Projekten) wird anhand etablierter Methoden der evidenzbasierten Medizin, Epidemiologie und Versorgungsforschung durchgeführt.</p> <p>Versorgungswege und Inanspruchnahme von medizinischen Versorgungsleistungen werden deskriptiv anhand von absoluten und relativen Häufigkeiten dargestellt, wobei die Maßgaben der Guten Praxis Sekundärdatenanalyse zur internen Diagnosevalidierung angewandt werden. Methoden der Alters- und Geschlechtsadjustierung und -standardisierung kommen zum Einsatz. Zur Beantwortung versorgungsrelevanter Fragestellungen werden sowohl Querschnittsanalysen als auch longitudinale Verfahren genutzt.</p> <p>Gesundheitsökonomische Analysen betrachten weiterhin direkte Kosten und indirekte Erkrankungskosten aufgrund Arbeitsunfähigkeit und Pflegeleistungen als Outcomes.</p> <p>Grundlage für sämtliche statistische Analysen stellt der jeweilige statistische Analyseplan dar, welcher durch einen befugten Projektmitarbeiter:in anhand der entsprechenden Handlungsanweisung vor Beginn der Datenanalyse zu erstellen und mit allen im Projekt beteiligten Mitarbeiter:innen, insbesondere den Projektleitern abzustimmen ist.</p>
<p>Ethische und rechtliche Aspekte</p>	<p>Das Projekt erfolgt im Einklang mit der Guten Epidemiologischen Praxis (GEP), der Guten Praxis Datenlinkage (GPD), den Bestimmungen der Berufsordnung für Ärzte und der Deklaration von Helsinki. Alle Datenschutzrechtlichen und regulatorischen Vorgaben werden eingehalten.</p> <p>Die Berichterstattung erfolgt unabhängig vom Projektergebnis nach den aktuellen Berichtsstandards für Sekundärdatenanalysen.</p>

Reference List

1. Schons MJ, Pilgram L, Reese, JP et al. (2022) The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. PREPRINT (Version 1) available at Research Square <https://doi.org/10.21203/rs.3.rs-1249111/v1>
2. Horn A, Krist L, Lieb W, et al.(2021) Long-term health sequelae and quality of life at least 6 months after infection with SARS-CoV-2: design and rationale of the COVIDOM-study as part of the NAPKON population-based cohort platform (POP). *Infection* 49(6) 1277-1287
3. Swart, E. and Ihle, P. (2005) Routinedaten im Gesundheitswesen - Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. Bern: Verlag Hans Huber.
4. Trautmann F, Meier F, Seidler A, Schmitt J (2016) Effects of the German skin cancer screening program on melanoma incidence and indicators of disease severity. *Br J Dermatol* .
5. Trautmann F, Schuler M, Schmitt J (2015) Burden of soft-tissue and bone sarcoma in routine care: Estimation of incidence, prevalence and survival for health services research. *Cancer Epidemiol* 39: 440-446.
6. Schmitt J, Romanos M, Schmitt NM, Meurer M, Kirch W (2009) Atopic eczema and attention-deficit/hyperactivity disorder in a population-based sample of children and adolescents. *JAMA* 301: 724-726.
7. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) (2005) Gute Praxis Sekundärdatenanalyse (GPS). In: Swart E, Ihle P, editors. Routinedaten im Gesundheitswesen - Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. Bern: Verlag Hans Huber. pp. 405-412.
8. Swart E, Bitzer EM, Gothe H, Harling M, Hoffmann F, Horenkamp-Sonntag D, Maier B, March S, Petzold T, Rohrig R, Rommel A, Schink T, Wagner C, Wobbe S, Schmitt J (2016) (A Consensus German Reporting Standard for Secondary Data Analyses, Version 2 (STROSA-STandardisierte BerichtsROutine für SekundärdatenAnalysen)) . *Gesundheitswesen* .
9. Roessler M, Tesch F, Batram M, et al. Post COVID-19 in children, adolescents, and adults: results of a matched cohort study including more than 150,000 individuals with COVID-19 (Preprint). *medRxiv* 2021.10.21.21265133; doi: <https://doi.org/10.1101/2021.10.21.21265133>

8.2 Anlage 1b: Probandeninformation (NAPKON – POP Kohortenstudie)

NAPKON – Nationales Pandemie Kohorten Netz

Proband:inneninformation

Populationskohorte COVIDOM:

Folgemorbidität von SARS-CoV-2-Infektionen – Auswirkungen auf Gesundheitszustand und Lebensqualität

Studienverantwortliche:

Prof. Dr. med. Stefan Störk Deutsches Zentrum für Herzinsuffizienz Medizinische Klinik und Poliklinik I Universitätsklinikum Am Schwarzenberg 15 97078 Würzburg	Prof. Dr. med. Peter U. Heuschmann, MPH Institut für Klinische Epidemiologie und Biometrie (IKE-B) Universität Würzburg Josef-Schneider-Straße 2 97080 Würzburg
Prof. Dr. Stefan Schreiber Universitätsklinikum Schleswig-Holstein (UKSH) Klinik für Innere Medizin I Arnold-Heller-Straße 3 Haus C, Haus K3 24105 Kiel	Prof. Dr. Heyo K. Kroemer Charité - Universitätsmedizin Berlin (Kordinierungsstelle Netzwerk Universitätsmedizin) Charitéplatz 1 10117 Berlin

Sehr geehrte:r Proband:in,

wir danken Ihnen, dass Sie diese Probanden:inneninformation lesen. Bitte überlegen Sie danach in aller Ruhe, ob Sie an der Populationskohorte „**COVIDOM: Folgemorbidität von SARS-CoV-2-Infektionen – Auswirkungen auf Gesundheitszustand und Lebensqualität**“ teilnehmen und dafür Ihre Daten und Biomaterialien (z. B. Blut, Urin, Gewebe) zur Verfügung stellen möchten.

Ihre Teilnahme an der Studie ist freiwillig und setzt Ihre schriftliche Einwilligung voraus. Wenn Sie nicht einwilligen oder Ihre Einwilligung später widerrufen, entstehen Ihnen daraus keine Nachteile.

Mit Fragen zu COVIDOM wenden Sie sich bitte an **Frau Anna Horn (Telefon: 0931/201- 47351; E-Mail: COVIDOM@ukw.de)**.

1. Hintergrund und Ziele der Studie

COVID-19 ist eine neuartige Erkrankung, die vom Virus SARS-CoV-2 ausgelöst wird. COVID-19 betrifft hauptsächlich die Lunge. Das Virus kann aber auch andere Organe auf bisher unbekannte Weise schädigen und bei einem Teil der Infizierten zu chronischen Folgeschäden führen. Um die Anstrengungen zur Erforschung der Infektion mit SARS-CoV-2 zu bündeln, wurde in Deutschland das Netzwerk Universitätsmedizin (NUM) gegründet. Das NUM wird durch die Charité – Universitätsmedizin Berlin zentral koordiniert (<https://www.netzwerk-universitaetsmedizin.de/impressum>). Ein Teil davon ist die Nationale Kohorten-Plattform (NAPKON). Auch die Studie COVIDOM wird mit Unterstützung des NUM durchgeführt.

In COVIDOM möchten wir mit SARS-CoV-2 Infizierte aus Unterfranken langfristig begleiten und nachuntersuchen. Dafür sind regelmäßige telefonische Befragungen (oder wunschweise das Ausfüllen von Online-Fragebögen) und mindestens ein längerer Untersuchungstermin am **Studienzentrum Würzburg** (bestehend aus der „**Survey Unit**“ des **DZHI** (Deutsches Zentrum für Herzinsuffizienz, Universitätsklinikum Würzburg (UKW), Am Schwarzenberg 15, 97078 Würzburg) und dem **IKE-B** (Institut für Klinische Epidemiologie und Biometrie der Universität, Josef-Schneider-Str. 2 / Haus D7 97080 Würzburg)) vorgesehen.

Mit der Studie möchten wir herausfinden

- welche Folgeerkrankungen eine SARS-CoV-2 Infektion verursacht und
- wie sich die aktuelle Pandemie auf unsere Gesellschaft auswirkt.

Dieses Wissen soll helfen, Personen mit einer SARS-CoV-2-Infektion oder einer COVID-19 Erkrankung besser zu behandeln und auf zukünftige Pandemien mit dem gleichen oder einem anderen Virus besser vorbereitet zu sein.

Es folgen nun Informationen zum Studienablauf. Wir möchten Sie an dieser Stelle darauf hinweisen, dass es Ihnen jederzeit möglich ist, auf einzelne der genannten Untersuchungen zu verzichten.

Was werden wir Sie fragen?

Wir möchten Sie während der Studie regelmäßig zu Ihrem Gesundheitszustand befragen, entweder in Form von Telefoninterviews oder mit Hilfe von Online-Fragebögen, die Sie auf Wunsch selbst ausfüllen können. Dabei werden Ihnen hauptsächlich Fragen zu Anzeichen einer Erkrankung, Ihrer Lebensqualität und zu Ihrer allgemeinen Lebenssituation gestellt.

Sofern bereits Krankenunterlagen in unserer Klinik von Ihnen vorliegen, würden wir gerne bereits vorhandene Angaben zu Ihrem Krankheitsverlauf daraus entnehmen.

Welche Untersuchungen sind geplant?

Bei der ersten Untersuchung am Studienzentrum Würzburg (ca. 6 bis 9 Monate nach Ihrer Infektion mit SARS-CoV-2) werden Ihre Lungen- und Herzfunktion sowie Ihre Bauchorgane untersucht. Hierzu werden auch bildgebende Verfahren, z. B.

Ultraschall-Untersuchungen durchgeführt. Außerdem sind eine Messung Ihrer Muskelkraft, Untersuchungen Ihrer Nerven und Augen, ein Riech- und Geschmackstest, eine Blutdruck-Messung und eine Blutabnahme geplant.

2. Welche Biomaterialien werden gesammelt und analysiert?

In COVIDOM sind mehrere Blutuntersuchungen geplant, die am Untersuchungstag oder am Folgetag erfolgen und deren Ergebnisse wir Ihnen unmittelbar im Anschluss mitteilen.

Darüber hinaus bitten wir Sie darum, ca. 35 ml zusätzliches Blut und einige weitere biologische Proben (z. B. Urin, Rachenabstrich) für die medizinische Forschung entnehmen zu dürfen. Diese Biomaterialien werden in den kommenden Monaten und Jahren durch Forscher:innen des Studienzentrums Würzburg und des NUM vielfältig wissenschaftlich untersucht.

Ihre Biomaterialien werden am Klinikum in der **zentralen Biobank** der Medizinischen Fakultät Würzburg, der **ibdw** (Interdisziplinäre Biomaterial- und Datenbank Würzburg) gelagert. Die zentrale Biobank unterliegt strengen Qualitäts- und Sicherheitsmaßnahmen. Ihre Bioproben werden dort bis zu 30 Jahre aufbewahrt. In besonderen Fällen können Daten und Bioproben auch über diesen Zeitraum hinaus von erheblicher Bedeutung für die Wissenschaft sein. In diesen Fällen würden wir in Abstimmung mit dem zuständigen Datenschutzbeauftragten und der unabhängigen Ethikkommission der Universität Würzburg klären, ob auch eine weitergehende Nutzung Ihrer Daten und Biomaterialien möglich ist.

3. Um welche Art von Daten bitten wir Sie?

Die Daten, die durch die Analyse Ihrer Bioproben und durch Ihre Untersuchungen entstehen, werden in einer Datenbank am Studienzentrum Würzburg und der zentralen Datenbank des NUM (siehe Punkt 8.) langfristig gespeichert.

Daneben bitten wir Sie um Ihre Erlaubnis, bei Bedarf folgende zusätzliche Daten über Sie anzufordern und wissenschaftlich zu nutzen:

Kontakt zu anderen Ärzt:innen

Eine wichtige Frage, die im Kontext der SARS-CoV-2 Infektion diskutiert wird, ist das Auftreten von möglichen Folgeerscheinungen und Langzeitfolgen durch die Infektion. Dabei ist aus heutiger Sicht nicht absehbar, wie sich solche Ereignisse äußern könnten und auf welchem Fachgebiet damit zu rechnen ist. Informationen von Haus- und Fachärzt:innen helfen dabei, Hinweise auf eine Verschlechterung möglicherweise schon zuvor bestehender Leiden festzustellen.

Wir bitten Sie daher darum, dass wir von Ihnen auch Daten, z. B. über vorangegangene und nachfolgende Ärzt:innenkontakte bei ambulanten Haus- und Fachärzt:innen sowie ggf. von weiteren Krankenhausaufenthalten und Arzneimittel-Verordnungen anfordern und wissenschaftlich nutzen dürfen.

In der Einwilligungserklärung können Sie uns dazu ermächtigen, die entsprechenden Daten innerhalb der nächsten 5 Jahre für die Nutzung in spezifischen, von der zuständigen Ethikkommission geprüften, wissenschaftlichen Projekten, anzufordern.

Nutzung weiterer Gesundheits- und Sozialdaten

Gesundheitsdaten sind Daten aus Ihrer Krankenversicherung. Sozialdaten sind Daten, die der gesetzlichen Rentenversicherung vorliegen. Innerhalb von NAPKON möchten wir diese Daten wissenschaftlich verwenden, um gesundheitsfördernde, aber auch schädliche Einflüsse auf Ihren Gesundheitsverlauf nach der SARS-CoV-2 Infektion festzustellen, die nicht mit medizinischen Tests oder telefonischen Befragungen nach ihrem Wohlbefinden erfassbar sind.

Durch die Ergänzung der Gesundheits- und Sozialdaten sollen folgende Fragen beantwortet werden:

1. Welcher Krankheitsverlauf zieht eine längere Arbeitsunfähigkeit nach sich?
2. Welcher Krankheitsverlauf erhöht die Wahrscheinlichkeit einer vorzeitigen Berentung?
3. Welche gemeinsamen Vorerkrankungen haben Personen die an Langzeitfolgen der COVID-19 Erkrankung leiden?
4. Welche ärztlichen Leistungen und Arzneimittel erhalten die Menschen im Krankheitsverlauf?
5. Welche Auswirkungen haben Maßnahmen der medizinischen Rehabilitation auf die Gesundheit und auf die Teilhabe am Arbeitsleben?

Zusätzlich bitten wir Sie um Ihre Erlaubnis, Gesundheits- und Sozialdaten anzufordern, um diese wissenschaftlich zu nutzen. Die Auswahl der Daten beschränkt sich auf den Beobachtungszeitraum 2 Jahre vor und 5 Jahre nach Ihrer Infektion und umfasst:

1. Daten aus Ihrer **gesetzlichen und/oder privaten Krankenversicherung**, die für die medizinische Versorgung im Rahmen Ihrer SARS-CoV-2 Infektion relevant sind:
 - Informationen zur ambulanten und stationären Versorgung (Diagnosen, erhaltene Leistungen, Verordnungen)
 - Heil- und Hilfsmittelverordnungen (nicht-ärztliche Leistungen wie Physiotherapie, Rollstühle etc.)
 - bei Erwerbstätigen Angaben zur Arbeitsunfähigkeit (Dauer, Diagnosen)

Hierfür werden wir Sie um Mitteilung Ihrer Krankenversicherung einschließlich Ihrer Krankenversicherungsnummer bitten.

2. Daten aus Ihrer **gesetzlichen Rentenversicherung**, die für die medizinische Rehabilitation im Rahmen Ihrer SARS-CoV-2 Infektion relevant sind und bei Ihrem gesetzlichen Rentenversicherungsträger bekannt sind:
 - medizinische und berufliche Rehabilitationsmaßnahmen
 - Arbeitslosigkeit
 - Rentenstatus

Hierfür werden wir Sie um Mitteilung Ihrer Rentenversicherungsnummer bitten. Ihren/Ihre Arbeitgeber:in werden wir nicht kontaktieren.

3. Daten, die von der **Kassenärztlichen Vereinigung oder den Gesundheitsämtern** zum Verlauf Ihrer SARS-CoV-2 Infektion gesammelt wurden.
4. Ggf. Daten aus **weiteren Register** in denen versorgungsrelevante Behandlungsdaten über mich gespeichert sind (z. B. Krebsregister)

4. Wofür werden Ihre Daten und Biomaterialien verwendet?

Da bislang noch sehr wenig über die Folgen einer SARS-CoV-2-Infektion bekannt ist, möchten wir Sie um eine **sehr umfassende Erlaubnis zur Verwendung Ihrer Biomaterialien und Daten** bitten. Über jede Nutzung durch Mitarbeiter des Studienzentrum Würzburgs wird die unabhängige Ethikkommission der Universität Würzburg vorher informiert und berät die Forscher:innen hierzu. Werden Ihre Daten und Bioproben in Projekten des NUM verwendet, die nicht von Forscher:innen des UKW geleitet werden, so übernimmt eine andere unabhängige Ethikkommission in Deutschland diese Aufgabe. Für diese national übergreifenden Forschungsvorhaben wird vom NUM außerdem ein Use and Access Komitee eingesetzt, das den wissenschaftlichen Nutzen jedes Forschungsprojektes vorab berät.

Das und die Einbindung unabhängiger Ethikkommissionen stellt sicher, dass Ihre Daten und Bioproben ausschließlich für sinnvolle und ethisch unbedenkliche Zwecke verwendet werden.

Ihre Biomaterialien und Daten sollen für eine ganze Reihe medizinischer Forschungsprojekte genutzt werden. Weil sich in der Forschung immer wieder neue Fragen ergeben, werden Ihre Biomaterialien und Daten wahrscheinlich auch für Projekte verwendet, deren Inhalt man heute noch nicht kennt. Alle Projekte verfolgen jedoch das Ziel, die **Vorbeugung, Erkennung und Behandlung von Erkrankungen** zu verbessern, die mit einer **SARS-CoV-2-Infektion oder möglichen zukünftigen Pandemien** in Verbindung stehen.

Da es Hinweise gibt, dass erbliche Faktoren den Verlauf einer SARS-CoV-2-Infektion beeinflussen können, werden sehr wahrscheinlich an Ihren Biomaterialien auch **genetische Untersuchungen** durchgeführt, unter Umständen auch eine **Untersuchung der gesamten darin enthaltenen Erbinformation**.

Das durch COVIDOM und andere Projekte des NUM entstehende Wissen soll auch auf andere Erkrankungen übertragen werden. Daher bitten wir Sie um Ihre Erlaubnis, mit Ihren Biomaterialien und Daten **alle denkbaren Auswirkungen einer SARS-CoV-2-Infektion** zu untersuchen, also z. B. auch Krebs, entzündliche Erkrankungen, Herz-Kreislauf-Erkrankungen und Erkrankungen des Gehirns.

Aus organisatorischen Gründen ist es nicht möglich, dass Sie die **Nutzung Ihrer Bioproben und Daten jetzt oder zu einem späteren Zeitpunkt näher eingrenzen** (z. B. durch den Ausschluss bestimmter Forschungsfragen oder ein Verbot der Weitergabe Ihrer Bioproben und Daten an Dritte). Wenn Sie mit der hier beschriebenen Art und Dauer der Nutzung nicht in vollem Umfang einverstanden sind, sollten Sie Ihre Einwilligung nicht erteilen.

5. Welche Risiken sind mit Ihrer Studienteilnahme verbunden?

a. Gesundheitliche Risiken (Blutentnahme)

Wir werden Ihnen nicht mehr als 60 ml Blut abnehmen (das entspricht ca. 4 Esslöffeln). Dies ist für Sie mit den geringen Risiken einer normalen Blutentnahme verbunden, d. h. die Einstichstelle kann schmerzen, oder es kann dort ein Bluterguss (blauer Fleck) entstehen. In äußerst seltenen Fällen kann sich ein Blutgerinnsel bilden (Thrombose), eine örtlich begrenzte Entzündung an der Einstichstelle auftreten, oder es kann zu dauerhaften Schädigungen von Blutgefäßen oder Nerven kommen.

Die Sammlung von Urinproben sowie die Durchführung eines Rachenabstrichs sind mit keinen gesundheitlichen Risiken für Sie verbunden.

b. Gesundheitliche Risiken (Untersuchungen)

Im Einzelfall ist es möglich, dass ein Auswertungsergebnis im Rahmen der COVIDOM-Studie für Ihre Gesundheit von so erheblicher Bedeutung ist, dass ein/e Ärzt*in oder Forscher*in eine Kontaktaufnahme als dringend notwendig erachtet. Das ist insbesondere der Fall, wenn sich daraus ein dringender Verdacht auf eine schwerwiegende, bisher möglicherweise nicht erkannte Krankheit ergibt, die behandelt oder deren Ausbruch verhindert werden könnte.

Darüber hinaus können sich weitere Analyseergebnisse ergeben, die möglicherweise für Ihre Gesundheit relevant sind (sog. Zusatzfunde) und über die wir Sie informieren möchten. Sie können entscheiden, ob wir Sie in diesem Zusammenhang kontaktieren dürfen. Beachten Sie dabei, dass Sie Gesundheitsinformationen, die Sie durch eine solche Rückmeldung erhalten, unter Umständen auch bei anderen Stellen (z. B. vor Abschluss einer Kranken- oder Lebensversicherung) offenbaren müssen und dadurch Nachteile erleiden könnten. Sie können die Kontaktaufnahmen in der Einwilligungserklärung ablehnen („Recht auf Nichtwissen“). Da auch genetische Untersuchungen möglich sind, weisen wir darauf hin, dass die hierbei gewonnenen genetischen Daten auch Bedeutung für Ihre Familienangehörigen haben können.

c. Weitere Risiken (Privatsphäre, Datenschutz)

Jede Verarbeitung, Speicherung und Übermittlung von Daten ist mit Vertraulichkeitsrisiken verbunden, d. h. Ihre Daten könnten unerlaubt wieder Ihrer Person zugeordnet werden. Auch in der Forschung lassen sich diese Risiken nicht völlig ausschließen und bestehen insbesondere dann, wenn Sie selbst umfangreiche Informationen über sich öffentlich preisgeben und z. B. zur Ahnenforschung genetische Daten im Internet oder in sozialen Netzwerke veröffentlichen. Unter **Punkt 8** („Wer hat Zugang zu Ihren Biomaterialien und Daten?“) erläutern wir Ihnen näher, wie Ihre Privatsphäre geschützt wird.

6. Welcher Nutzen ergibt sich für Sie persönlich?

Das Untersuchungsprogramm von COVIDOM kombiniert gängige klinische Untersuchungen mit Untersuchungen, die allein der Forschung dienen und deren Ergebnisse noch keine klinische Anwendung haben. Zu den gängigen klinischen Untersuchungen bekommen Sie im Anschluss einen **Ergebnisbrief**, der unter anderem die Resultate der Messungen von Lungenfunktion und Herzfunktion enthält. Auch die Ergebnisse der zeitnahen Laboruntersuchungen werden wir Ihnen darin mitteilen. **Die reinen Forschungsdaten ohne klinische Verwendung werden Ihnen jedoch in der Regel nicht mitgeteilt.**

7. Welcher Nutzen ergibt sich für unsere Gesellschaft?

Die SARS-CoV-2-Pandemie hat schwere gesundheitliche, wirtschaftliche und soziale Folgen. Es gibt daher ein großes gesellschaftliches Interesse daran, die Auswirkungen der Pandemie auf allen Ebenen besser zu verstehen. Daher erwarten wir, dass COVIDOM, die Kohorten des NAPKON und alle anderen Projekte des NUM einen großen Nutzen für die Allgemeinheit haben werden.

8. Wer hat Zugang zu Ihren Bioproben und Daten, und wie werden diese geschützt?

a. Kennzeichnung Ihrer Bioproben und Daten

Die lokale Organisation, Datenerhebung, das Datenmanagement und die Datenauswertung am Studienzentrum Würzburg erfolgen durch das DZHI (Prof. Dr. S. Störk, Universitätsklinikum Würzburg) und das IKE-B (Prof. Dr. P. Heuschmann, Universität Würzburg). Alle Ihre identifizierenden Daten (Name, Geburtsdatum, Anschrift) werden unverzüglich nach der Erhebung Ihrer medizinischen Daten bzw. der Gewinnung Ihrer Bioproben von diesen getrennt und durch einen Code ersetzt („pseudonymisiert“). Nach einer weiteren Umkodierung werden die medizinischen Daten gespeichert und die Bioproben lokal eingelagert.

Der zentrale Speicherort Ihrer medizinischen Daten ist die Studiendatenbank am UKSH, ihre Bioproben werden in der Biomaterial- und Datenbank der Medizinischen Fakultät Würzburg (ibdw) gelagert. Die Pseudonymisierung schließt eine unbefugte Zuordnung der Daten und Biomaterialien zu Ihrer Person weitestgehend aus. Biomaterialien und Daten werden nur in **doppelt kodierter** („doppelt pseudonymisierter“) Form vom Studienzentrum Würzburg/der ibdw bzw. dem NUM für Forschungszwecke herausgegeben.

Ihre pseudonymisierten medizinischen Daten werden ebenfalls an die zentrale Datenbank des NUM (Charité Universitätsmedizin Berlin) übermittelt. Die Speicherung in der NUM_Datenbank erfolgt unter einem anderen Pseudonym. Die bei uns gespeicherten und an das NUM übermittelten Bilddaten werden in der Datenbank des NUM ebenfalls pseudonymisiert gespeichert. Für die pseudonymisierte Speicherung von Bilddaten stellt das NUM eine eigene Datenbank zur Verfügung. Um die Verwaltung dieser Daten zu ermöglichen, wird eine **unabhängige Treuhandstelle** (Unabhängige Treuhandstelle der Universitätsmedizin Greifswald, Ellernholzstr. 1-2, 17475 Greifswald; <https://www.ths-greifswald.de/kontakt>) die elektronische Verwaltung der Einwilligungsunterlagen für das NUM übernehmen und die medizinischen Daten für die zentralen Datenbanken pseudonymisieren. **Dafür werden Ihre identifizierenden Daten (Name, Geburtsdatum, Anschrift) an die Treuhandstelle übermittelt.**

Ohne die Mitwirkung der Treuhandstelle können Ihre Daten und Bioproben im NUM-Netzwerk nicht mehr Ihrer Person zugeordnet werden. Eine solche Zuordnung erfolgt nur, wenn zusätzliche Daten aus Ihren Krankenunterlagen erhoben werden sollen oder wir erneut mit Ihnen in Kontakt treten möchten. Ihre identifizierenden Daten werden nicht an Forscher*innen oder unberechtigte Dritte, wie z. B. Versicherungen oder Arbeitgeber, weitergegeben.

b. Weitergabe von Daten und Bioproben

Ihre medizinischen Daten und Bioproben können vom Studienzentrum Würzburg/der Biomaterial- und Datenbank der Medizinischen Fakultät Würzburg (ibdw) und vom NUM für bestimmte medizinische Forschungszwecke in pseudonymisierter Form an andere Einrichtungen z. B. Universitäten, Forschungsinstitute und forschende Unternehmen **innerhalb der EU** weitergegeben werden. Unter Umständen werden Ihre Daten und Bioproben dabei auch mit Daten aus anderen Quellen verknüpft, sofern die gesetzlichen Voraussetzungen hierfür erfüllt sind. Ihre Daten und Bioproben dürfen in jedem Fall nur für den vorbestimmten Forschungszweck verwendet und vom Empfänger nicht unerlaubt weitergegeben werden. Nicht verbrauchtes Biomaterial wird an die Biomaterial- und Datenbank der Medizinischen Fakultät Würzburg (ibdw) zurückgegeben oder vernichtet.

Ihre Daten und Bioproben können auch an Empfänger in Ländern **außerhalb der EU** weitergegeben werden, wenn die Europäische Kommission bei dem entsprechenden Land ein angemessenes Datenschutzniveau festgestellt oder das Studienzentrum Würzburg bzw. das NUM mit den jeweiligen Forschungspartnern verbindlich die Gewährleistung eines angemessenen Daten-Schutzniveaus vereinbart haben.

c. Weitergabe in Drittländer außerhalb der EU

Darüber hinaus kann es aber auch vorkommen, dass Daten und Bioproben an Forschungspartner in Drittländern weitergegeben werden sollen, für die **keine** dieser beiden Voraussetzungen erfüllt ist. Diese Länder haben möglicherweise ein niedrigeres Datenschutzniveau als die EU. COVIDOM sichert zu, auch in diesen Fällen die Forschungspartner vertraglich, soweit rechtlich möglich, zur Einhaltung des EU-Datenschutzniveaus zu verpflichten. Dennoch besteht das Risiko, dass staatliche oder private Stellen auf Ihre Daten zugreifen, obwohl dies nach dem europäischen Datenschutzrecht nicht zulässig wäre. Zudem kann es sein, dass Ihnen dort weniger oder schlechter durchsetzbare Betroffenenrechte zustehen und es keine unabhängige Aufsichtsbehörde gibt, die Sie bei der Wahrnehmung ihrer Rechte unterstützen könnte. **Eine Weitergabe Ihrer Daten und Proben kann in diesem Fall nur erfolgen, wenn Sie dem ausdrücklich zugestimmt haben. Dazu können Sie in der Einwilligungserklärung das entsprechende Kästchen ankreuzen.**

d. Bewertung durch die Ethik-Kommission

Grundsätzliche Voraussetzung für die Nutzung Ihrer Daten und Bioproben für ein konkretes Forschungsprojekt ist die vorherige positive Bewertung des Forschungsprojekts durch eine unabhängige Ethikkommission.

e. Veröffentlichungen

Veröffentlichungen von Forschungsergebnissen erfolgen ausschließlich in anonymisierter Form, so dass sie keine Rückschlüsse auf Ihre Person zulassen. Das gilt insbesondere für die Veröffentlichung genetischer Forschungsergebnisse. Allerdings verlangen die Herausgeber wissenschaftlicher Fachzeitschriften unter Umständen die Hinterlegung der für die Forschung genutzten Daten in Datenbanken, die jedoch besonders geschützt und nicht für die Allgemeinheit zugänglich sind.

9. Hat jemand finanzielle Vorteile von der Nutzung Ihrer Daten und Bioproben?

Für die Überlassung Ihrer Daten und Bioproben erhalten Sie kein Entgelt. Sollte aus der Forschung an Ihren Daten und Bioproben ein kommerzieller Nutzen entstehen, so werden Sie daran nicht beteiligt. Mit der Überlassung der Bioproben an das Studienzentrum Würzburg/ UKW werden die Bioproben zu **1/3 Eigentum des Studienzentrums Würzburg/UKW** und zu **2/3 Eigentum des NUM**. Mit der Eigentumsübertragung räumen Sie dem Studienzentrum Würzburg/UKW und dem NUM die **Nutzungsrechte an Ihren Daten und Bioproben** ein. Ihre Daten und Bioproben werden nicht verkauft. Das Studienzentrum Würzburg/UKW und das NUM können jedoch von Nutzern Ihrer Daten und Bioproben eine angemessene Aufwandsentschädigung erheben.

10. Erneute Kontaktaufnahme

Im Rahmen der COVIDOM-Studie möchten wir Sie mindestens ein Jahr lang regelmäßig (ca. **alle 6 bis 12 Monate**) kontaktieren und zu Ihrem Gesundheitsstatus befragen. Abhängig von den Ergebnissen der ersten Untersuchungstermine am Studienzentrum Würzburg sind darüber hinaus regelmäßige Nachuntersuchungen möglich. Ausnahmsweise kann es auch vorkommen, dass wir Sie kontaktieren, um weitere Daten von Ihnen zu erheben bzw. weitere Bioproben zu gewinnen, um Sie um die Erlaubnis zu bitten, Ihre Daten aus der COVIDOM-Studie mit Daten aus anderen Quellen zu verknüpfen, oder aber um Sie über weitere Forschungsprojekte des Nationalen Forschungsnetzwerks Universitätsmedizin zu informieren, wenn Sie damit einverstanden sind.

11. Was beinhaltet Ihr Widerrufsrecht?

Sie können Ihre Einwilligung in die Teilnahme an der COVIDOM-Studie im Rahmen von NAPKON und in die Nutzung Ihrer Daten und Bioproben jederzeit ohne Angabe von Gründen und ohne persönliche Nachteile widerrufen.

Die Rechtmäßigkeit der bis zum Widerruf erfolgten Nutzung Ihrer Daten und Bioproben bleibt davon unberührt. Im Falle eines Widerrufs werden Ihre Bioproben vernichtet, außer Sie entscheiden, dass die Proben in anonymisierter Form weiter für Forschungszwecke verwendet werden dürfen. Die Anonymisierung kann eine spätere Zuordnung von Bioproben zu Ihrer Person durch einen genetischen Abgleich mit anderen Quellen jedoch nicht vollständig ausschließen.

Im Falle eines Widerrufs können Sie außerdem entscheiden, ob die von Ihnen bereits erhobenen Daten gelöscht werden sollen oder in anonymisierter Form weiterverwendet werden dürfen. Die Datenlöschung erfolgt im Rahmen der technischen Möglichkeiten und nur soweit sie die Verwirklichung der Ziele von **NAPKON** und des **NUM** dadurch nicht unmöglich macht oder unangemessen beeinträchtigt.

Nachdem Ihre Daten und Bioproben anonymisiert wurden, ist eine anschließende Löschung bzw. Vernichtung nicht mehr möglich. Daten aus bereits durchgeführten oder bereits publizierten Analysen können nicht mehr gelöscht werden.

Für einen Widerruf Ihrer Einwilligung und die Ausübung Ihrer Rechte wenden Sie sich bitte an Ihr zuständiges Studienzentrum:

Studienzentrum Würzburg, Deutsches Zentrum für Herzinsuffizienz, Haus A15,
Am Schwarzenberg 15, 97078 Würzburg.

Telefon: 0931/201- 47300 oder E-Mail: COVIDOM@ukw.de

12. Welche weiteren Rechte haben Sie?

Rechtsgrundlage für die Verarbeitung Ihrer Daten ist Ihre Einwilligung (Art. 6 Abs. 1 Buchst. a und Art. 9 Abs. 2 Buchst. a der Europäischen Datenschutz-Grundverordnung (DSGVO)).

Verantwortliche für die **Datenverarbeitung** im Sinne der DSGVO sind:

- **das UKW**, Anstalt des öffentlichen Rechts, Josef-Schneider-Straße 2, 97080 Würzburg, Telefon: 0931-201-0, oder E-Mail: info@ukw.de.

Das UKW hat einen behördlichen Datenschutzbeauftragten bestellt. An ihn können Sie sich mit allen Anliegen rund um Ihre Daten wenden oder auch mit einer Beschwerde über Datenschutzverstöße. Seine Kontaktdaten lauten wie folgt:

Datenschutzbeauftragter des UKW, Josef-Schneider-Straße 11, 97080 Würzburg, E-Mail: datenschutz@ukw.de, Telefon: +49 931 201-55485,

- **die Universität Würzburg**, vertreten durch das IKE-B, Vorstand Prof. Dr. P. U. Heuschmann, Josef-Schneider-Str. 2, Haus D7, 97080 Würzburg, E-Mail: epidemiologie@klinik.uni-wuerzburg.de; Telefon: +49 931 201-47308.
Auch die Universität Würzburg hat einen behördlichen Datenschutzbeauftragten bestellt. An ihn können Sie sich mit allen Anliegen rund um Ihre Daten wenden oder auch mit einer Beschwerde über Datenschutzverstöße. Seine Kontaktdaten lauten wie folgt:
Datenschutzbeauftragter der Universität Würzburg, Sanderring 2, 97070 Würzburg, E-Mail: datenschutz@uni-wuerzburg.de,

- **das Universitätsklinikum Schleswig-Holstein (UKSH)**, vertreten durch Prof. Dr. S. Schreiber Arnold-Heller-Straße 3 24105 Kiel Telefon: +49 431 500-0.
Auch das UKSH hat einen behördlichen Datenschutzbeauftragten bestellt. An ihn können Sie sich mit allen Anliegen rund um Ihre Daten wenden oder auch mit einer Beschwerde über Datenschutzverstöße. Seine Kontaktdaten lauten wie folgt:
Datenschutzbeauftragter UKSH: Dr. Stefan Reuschke, UKSH, Campus Lübeck, Maria-Goeppert-Str. 7b, 23538 Lübeck, Telefon: 0451/500-14180; E-Mail: Stefan.Reuschke@uksh.de, und
- **die Charité Universitätsmedizin Berlin**, Vorstandsvorsitzender Prof. Dr. H. K. Kroemer, als Studienzentrum und koordinierende Stelle des Nationalen Forschungsnetzwerks Universitätsmedizin (NUM) Charitéplatz 1 10117 Berlin, Telefon: +49 30 450 570 001.
Auch die Charité Universitätsmedizin Berlin hat einen behördlichen Datenschutzbeauftragten bestellt. An ihn können Sie sich mit allen Anliegen rund um Ihre Daten wenden oder auch mit einer Beschwerde über Datenschutzverstöße. Seine Kontaktdaten lauten wie folgt: **Datenschutzbeauftragten der Charité Universitätsmedizin Berlin** wenden: Charitéplatz 1, 10117 Berlin, Telefon: 030 450 580 016, E-Mail: datenschutz@charite.de.

Sie können vom UKW, von der Universität Würzburg, dem UKSH, der Charité Berlin und dem NUM im Rahmen der gesetzlichen Vorgaben eine Auskunft über die von Ihnen gespeicherten Daten verlangen. Sie können eine Berichtigung falscher Daten, eine Übertragung Ihrer Daten an Sie sowie eine Löschung der Daten oder eine Einschränkung ihrer Verarbeitung verlangen. Für die Ausübung dieser Rechte wenden Sie sich bitte an den Datenschutzbeauftragten des UKW oder der Universität Würzburg.

Die genannten Rechte können eingeschränkt werden, wenn deren Verwirklichung das Erreichen der Ziele von NAPKON (COVIDOM-Studie) und des NUM unmöglich macht oder unangemessen beeinträchtigt.

Sie haben außerdem ein Beschwerderecht bei **jeder** Datenschutzaufsichtsbehörde.

Die für das **UKW und die Universität Würzburg** zuständige Datenaufsichtsbehörde ist der **Bayerische Landesbeauftragte für den Datenschutz**, Postfach 22 12 19, 80502 München, E-Mail: poststelle@datenschutz-bayern.de.

Die zuständige Aufsichtsbehörde für die **Charité Universitätsmedizin Berlin** ist die **Berliner Beauftragte für Datenschutz und Informationsfreiheit**, Friedrichstr. 219, 10969 Berlin. Tel.: +49 30 13889-0, Fax: +49 30 2155050, E-Mail: mailbox@datenschutz-berlin.de.

Die zuständige Aufsichtsbehörde für das **Universitätsklinikum Schleswig-Holstein (UKSH)** ist: **Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein** Postfach 71 16 24171 Kiel Holstenstraße 98 24103 Kiel Telefon: 04 31/988-12 00 Telefax: 04 31/988-12 23 E-Mail: mail@datenschutzzentrum.de

Eine Liste der Aufsichtsbehörden in Deutschland finden Sie unter:

https://www.bfdi.bund.de/DE/Infothek/Anschriften/Links/anschriften_links-node.html.

13. Wo können Sie weitere Informationen erhalten?

Für die Beantwortung von weiteren Fragen zu dieser Aufklärung und zur COVIDOM-Studie im Rahmen des NAPKON-Projekts steht Ihnen die **Studienkoordinatorin**, Frau Anna Horn zur Verfügung (Telefon: 0931/201- 47351; E-Mail: COVIDOM@ukw.de).

8.3 Anlage 1c: Einwilligungserklärung (NAPKON – POP Kohortenstudie)

NAPKON – Nationales Pandemie Kohorten Netz Einwilligungserklärung

Populationskohorte COVIDOM:

Folgemorbidität von SARS-CoV-2-Infektionen – Auswirkungen auf Gesundheitszustand und Lebensqualität

Version 2.0 – 2022-01-28

Die grauen Felder sind vom Studienpersonal auszufüllen

<p>pheno – </p> <p>Codierung erfolgt nach Unterschrift</p>

Name, Vorname des/der Proband:in	<input type="checkbox"/> m <input type="checkbox"/> w	Geburtsdatum:
		Geburtsort:

Ich bin in einem persönlichen Gespräch ausführlich und verständlich über Wesen, Bedeutung, Risiken und Tragweite der Verarbeitung von Proband:innendaten zur medizinischen Forschung zur Erkennung, Behandlung und Vorbeugung von der SARS-CoV-2 Infektion und der COVID-19 Erkrankung aufgeklärt worden.

Ich habe darüber hinaus die Texte der Proband:inneninformation sowie die hier nachfolgend abgedruckte Einwilligungserklärung gelesen. Ich hatte die Gelegenheit, mit geschulten Studienmitarbeitern über die Durchführung der **COVIDOM**-Studie im Rahmen des nationalen NAPKON-Projekts zu sprechen. Alle meine Fragen wurden zufriedenstellend beantwortet.

Ich hatte ausreichend Zeit, mich zu entscheiden.

Datenschutz:

Ich erkläre mich damit einverstanden, dass das **Studienzentrum Würzburg** im Rahmen der NAPKON-Studie **COVIDOM** personenidentifizierende Daten und Proband:innendaten (einschließlich Bilddaten), wie in der Proband:inneninformation beschrieben, von mir erheben und speichern darf. Weiterhin erkläre ich mich damit einverstanden, dass zusätzliche Angaben zu meinem Krankheitsverlauf aus meinen Krankenunterlagen sowie den durchgeführten studienspezifischen Untersuchungen

und Befragungen entnommen und gespeichert werden dürfen. Ich erkläre mich damit einverstanden, dass meine Proband:innendaten **für 30 Jahre ab dem Zeitpunkt meiner Einwilligung im Studienzentrum Würzburg und der Datenbank des Nationalen Forschungsnetzwerks Universitätsmedizin - NUM gespeichert und verwendet werden.**

Meine Daten dürfen codiert zu Zwecken der Erforschung der SARS-CoV-2 Infektion und der COVID-19 Erkrankung (auch zur Qualitätssicherung) sowie zukünftiger Pandemien, wie in der Proband:inneninformation beschrieben, weitergegeben werden.

Ich willige ein, an der Biomaterialsammlung teilzunehmen. Ich bin damit einverstanden, dass meine Bioproben, wie in der Proband:inneninformation beschrieben, am Uniklinikum Würzburg (UKW) in der zentralen Interdisziplinäre Biomaterial- und Datenbank der Medizinischen Fakultät Würzburg (**ibdw**) qualitätskontrolliert gelagert werden.

Auch diese dürfen, wie in der Proband:inneninformation beschrieben, vom **Studienzentrum Würzburg/UKW** und über das **NUM** für wissenschaftliche Fragestellungen in einem wesentlichen Zusammenhang mit einer SARS-CoV-2 Infektion verwendet und zu diesem Zweck auch weitergegeben werden. Dies schließt unter Umständen auch die Weitergabe für Forschungsprojekte in Länder außerhalb der EU ein. Eine solche Weitergabe erfolgt nur, wenn für das Empfängerland sichergestellt werden kann, dass es über ein den Anforderungen der europäischen Datenschutzvorschriften entsprechendes Schutzniveau verfügt.

Darüber hinaus kann es aber auch vorkommen, dass meine Daten und Bioproben an Forschungspartner in Drittländern weitergegeben werden sollen, die möglicherweise ein niedrigeres Datenschutzniveau als die EU haben. COVIDOM sichert zu, auch in diesen Fällen die Forschungspartner vertraglich, soweit rechtlich möglich, zur Einhaltung des EU-Datenschutzniveaus zu verpflichten. Über dennoch mögliche Risiken, die bei einem Transfer meiner Daten und/oder Bioproben in solche Drittländer bestehen, wurde ich aufgeklärt (Proband:inneninformation Punkt 8.c). **Eine Weitergabe meiner Daten und Proben kann in diesem Fall nur erfolgen, wenn ich dem ausdrücklich zustimme. Dazu kann ich in der Tabelle auf Seite 3 das entsprechende Kästchen ankreuzen.**

Mir ist bewusst, dass bei jeder Erhebung, Speicherung und Übermittlung von Daten durch das Hinzuziehen weiterer Informationen, z. B. aus dem Internet oder sozialen Netzwerken, das Restrisiko einer Rückverfolgbarkeit zu meiner Person besteht. Das Risiko einer Rückverfolgbarkeit ist bei genetischen Proband:innendaten erhöht.

Ich bin damit einverstanden, dass ich im Rahmen der COVIDOM-Studie zu den angegebenen Zeitpunkten erneut kontaktiert werde.

Ich bin damit einverstanden, dass meine originalen medizinischen oder studienspezifischen Unterlagen im Studienzentrum Würzburg durch eine/n Beauftragte:n der COVIDOM-Studie im Rahmen eines so genannten Monitorings eingesehen werden können.

Ich bin darüber informiert, dass die Ergebnisse der Studie anonymisiert veröffentlicht werden.

Das Nutzungsrecht an meinen Daten übertrage ich an das **Würzburger Studienzentrum/ UKW** und **NUM**.

Mit der Überlassung meiner Bioproben an das Würzburger Studienzentrum/UKW werden diese zu **1/3 Eigentum des Würzburger Studienzentrums/UKW** und zu **2/3 Eigentum des NUM**. Mit der Eigentumsübertragung räume ich dem Würzburger Studienzentrum/UKW und dem NUM auch die Nutzungsrechte an den Bioproben ein.

Des Weiteren verzichte ich auf eventuelle urheber-, patent- und gebrauchsmusterrechtliche Ansprüche, soweit diese aus dem Erkenntnisgewinn eines NAPKON-Forschungsprojektes wie der COVIDOM-Studie resultieren.

Trotz Übertragung der Nutzungsrechte kann ich meine Einwilligung in die Datenverarbeitung jederzeit widerrufen.

Ich willige ein, dass im Rahmen dieser Studie persönliche Gesundheits- und Sozialdaten von mir von den nachfolgend genannten Institutionen abgefragt, gespeichert und zum Zweck der Studie genutzt werden dürfen – und entbinde diese für diesen Zweck von ihrer Schweigepflicht gegenüber meinem Studienzentrum zur Erfassung studienrelevanter Informationen. Diese Entbindung von der Schweigepflicht umfasst:	
<ul style="list-style-type: none"> • Haus- und Fachärzte sowie Krankenhäuser, bei denen ich in der Vergangenheit in Behandlung war 	<input type="checkbox"/> ja <input type="checkbox"/> nein
<ul style="list-style-type: none"> • meine Krankenversicherung 	<input type="checkbox"/> ja <input type="checkbox"/> nein
<ul style="list-style-type: none"> • die Rentenversicherung Bund bzw. meinen gesetzlichen Rentenversicherungsträger (s. Abschnitt 3 der Patientinnen und Patienteninformation – Basis-Modul) 	<input type="checkbox"/> ja <input type="checkbox"/> nein
<ul style="list-style-type: none"> • das Gesundheitsamt bei denen Daten über meine Infektion mit SARS-CoV-2 gespeichert sind 	<input type="checkbox"/> ja <input type="checkbox"/> nein
<ul style="list-style-type: none"> • weitere Register in denen versorgungsrelevante Behandlungsdaten über mich gespeichert sind (z. B. Krebsregister) 	<input type="checkbox"/> ja <input type="checkbox"/> nein

Darüber hinaus stimme ich der codierten Weitergabe meiner Daten und Bioproben in Länder außerhalb der EU auch in den Fällen zu, in denen kein dem europäischen Datenschutzrecht entsprechendes Schutzniveau sichergestellt werden kann. Über die möglichen Risiken einer solchen Weitergabe wurde ich aufgeklärt (Proband:innen-information Punkt 8.c).	<input type="checkbox"/> ja <input type="checkbox"/> nein
--	--

Ich bin damit einverstanden, (falls nicht gewünscht, bitte „nein“ ankreuzen)

dass ich zu einem späteren Zeitpunkt erneut kontaktiert werde,	
<ul style="list-style-type: none"> • um Informationen für zusätzliche wissenschaftliche Fragen zur Verfügung zu stellen (Proband:inneninformation Punkt 10) • um Informationen über weitere Forschungsprojekte des Nationalen Forschungsnetzwerks Universitätsmedizin (NUM) zu erhalten (Proband:inneninformation Punkt 10) • um meine Einwilligung in die Verknüpfung meiner Proband:innendaten mit medizinischen Informationen aus anderen Datenbanken einzuholen (Proband:inneninformation Punkte 8b und 10) 	<input type="checkbox"/> ja <input type="checkbox"/> nein

dass ich zu einem späteren Zeitpunkt erneut kontaktiert werde, <ul style="list-style-type: none"> zum Zweck der Rückmeldung medizinischer Zusatzfunde (Proband:innen-information Punkt 5b) 	<input type="checkbox"/> ja <input type="checkbox"/> nein
--	--

Eine solche Rückmeldung soll erfolgen über die Einrichtung, in der meine Daten gewonnen wurden oder über folgend genannte/n Ärzt:in: *(falls gewünscht, bitte angeben)*

Name und Anschrift der/des Ärztin:Arztes: Bitte vollständig und in Druckbuchstaben angeben!

Eine Kopie der Proband:inneninformation zur COVIDOM-Studie im Rahmen des NAPKON-Projekts und der Einwilligungserklärung Version 2.0 vom 28.22.2021 habe ich erhalten. Das Original verbleibt im **Studienzentrum Würzburg**.

Name, Vorname des/der Proband:in in Druckbuchstaben:	Ort:
Unterschrift des/der Proband:in:	Datum:

Ich habe das Aufklärungsgespräch geführt und die Einwilligung des/der Probandin:in zur Teilnahme an der **COVIDOM-Studie im Rahmen des NAPKON-Projekts** eingeholt.

Name, Vorname der/des Studienmitarbeiters:in in Druckbuchstaben:	Ort:
Unterschrift der/des Studienmitarbeiters:in:	Datum:

8.4 Anlage 2: FOSA – Befragung

FOSA Befragung Fachbeirat

Über eine anonyme Online-Umfrage, die über die im Netzwerk Universitätsmedizin (NUM) eingerichteten fächer- und organspezifischen Arbeitsgruppen (FOSAs) durchgeführt wurde, sollten die Meinungen und Standpunkte der Kliniker und klinischen Wissenschaftler in die Empfehlungen des Gutachtens einfließen. Zuvor wurden in einer Sitzung des NUM-FOSA-Beirats, in dem die Sprecher aller FOSA organisiert sind, der Gutachtauftrag sowie die in der Umsetzung befindliche Machbarkeitsstudie zur Verknüpfung von NAPKON-Daten mit Daten der gesetzlichen Krankenversicherungen vorgestellt. Der verwendete Fragebogen befindet im ANHANG.

Der verwendete Fragebogen ist am Ende des Dokuments. Der Fragebogen enthielt Fragen zur Person der "Befragten", Fragen zu den Erfahrungen mit der Nutzung von GKV-Daten, Fragen zur Datenverknüpfung, Fragen zu Barrieren bei der Datennutzung und zu Hilfsmöglichkeiten. Insgesamt wurden 14 Fragen gestellt, von denen 7 Freitextfragen waren.

Fragen 1-3:

Insgesamt wurden von 30 Befragten Fragebogen ausgefüllt. Etwa die Hälfte der 29 bestehenden NUM FOSAs waren durch Vertreter:innen (14/29) abgedeckt. Es lagen 28 auswertbare Fragebögen vor. Fehlende Werte stellen im Allgemeinen die Differenz zu dieser Gesamtzahl dar.

1. Geschlechterverteilung:

Weiblich: 14

Männlich: 14

2. Altersverteilung

<35 Jahre	5
36-40 Jahre	5
41-45 Jahre	3
46-50 Jahre	1
51-55 Jahre	10
56-60 Jahre	1
>60 Jahre	2

3. Welcher Fach- und Organspezifischer Arbeitsgruppe gehören Sie an?

1	Rehabilitation
2	Neurologie
3	Schmerzmedizin
4	Anästhesiologie & Intensivmedizin

5	Allgemeinmedizin
6	-/-
7	FOSA Allgemeinmedizin
8	Wissenschaftliche Mitarbeiterin
9	Epidemiologie und Public Health
10	FOSA Epidemiologie
11	Geriatrie
12	Physikalische und Rehabilitative Medizin
13	Epidemiologie & Public Health
14	FOSA Radiologie
15	Allgemeinmedizin, Versorgungsforschung
16	RACOON Combine
17	Akutschmerz, Universitätsklinikum
18	Arzt in Weiterbildung
19	Institut für Allgemeinmedizin, Versorgungsforschung
20	GKV-Routinedaten
21	Arbeitsgruppe Routinedaten, mit infektiologischem Schwerpunkt
22	Intensivmedizin Anästhesiologie
23	Neuroradiologie
24	Geriatrie
25	Allgemeinmedizin
26	Psychische Gesundheit
27	LokS
28	Laboratoriumsmedizin
29	Epidemiologie, Zahnmedizin
30	Kardiologie, Epi/PH

Frage 4: Befürworten Sie es, dass auf Basis der informierten Einwilligung durch Patientinnen und Patienten (Informed Consent) externe Daten (z. B. GKV-Routinedaten von Krankenkassen, Registerdaten und Befragungsdaten) individuell verknüpft und für die Gesundheitsforschung genutzt werden?

Ausnahmslos wurde von allen Antwortenden (n=26) befürwortet, dass auf der Basis einer informierten Einwilligung durch Patientinnen und Patienten externe Daten, zum Beispiel GKV Routinedaten von Krankenkassen, Registerdaten und Befragungsdaten, individuell verknüpft und für die Gesundheitsforschung genutzt werden.

Frage 5. Welche konkreten Forschungsfrage(n) halten Sie für besonders geeignet um sie mittels Verknüpfung mit GKV-Routinedaten von Krankenkassen im Kontext von NAPKON zu beforschen?

1	Weiterverfolgung von AU Zeiten, Nutzung ambulanter Hei- und Hilfsmittel und deren Einfluss auf Teilhabe und Aktivität Inanspruchnahme von Gesundheitsleistungen insgesamt
2	a) Langzeitverlauf Post-Covid Symptome verknüpfen mit AU-Zeiten, direkten und indirekten Krankheitskosten im Vergleich zu Propensity Score gematchten Kontrollen ohne COVID-19 b) Frühere Erkrankungen als Risikofaktoren für Post-Covid Symptome identifizieren

3	sektorenübergreifende Langzeitbeobachtungen
4	Hospitalisierung vgl. Kostenanalysen für ambulante Versorgung/Klinikversorgung (Kosten Versorgung, Maßnahmen, Therapien) Untersuchung des Versorgungspfades (welche Versorgungsstrukturen wurden in Anspruch genommen und wie häufig (längsschnittige Untersuchung Vergleichende Analysen mit gesunden Kohorten
5	Ermittlung von Risikofaktoren
6	ist die Versorgung von Patienten mit Osteoporose leitliniengerecht?
7	Langzeitoutcome, Inanspruchnahme von Ressourcen, Erfassung Einflussfaktoren vor Infektion
8	neurologische / psychiatrische Erkrankungen post COVID
9	Klinische Prädiktoren für hohe Leistungsanspruchnahme innerhalb von ein bis zwei Jahren nach Covid-Infektion
10	Langzeitfollowup
11	Palliativmedizinische Versorgung von VS mit COVID-19
12	Kosten-Nutzen-Effektivität von verschiedenen Therapien bei COVID-19; Inanspruchnahme Gesundheitssystem nach COVID-19 (inkl. Erkrankungen, die gehäuft nach COVID-19 auftreten)
13	Langzeitfolgen von COVID-19 (Krankheitslast, Folgeerkrankungen, Kosten); Long-COVID (Verlauf, Krankheitslast, Risikofaktoren, assoziierte Erkrankungen, Kosten)
14	LongCovid Langzeitoutcome und Arbeitsfähigkeit nach Intensivtherapie (PICS)
15	Schlaganfall Outcome. Outcome nach Aneurysma-Ruptur. Rupturrate von intrakraniellen Aneurysmata
16	Versorgungsaspekte wie Komorbidität, Behandlungsdauer, Behandlungshäufigkeit
17	Risikofaktoren basierte Fragestellungen Mortalitätsstatistiken seltene Erkrankungen (auch wenn Codierung zu wünschen übrig lässt) Medikamenten-Interaktionen
18	Prospektive Assoziation mit Erkrankungsendpunkten; Outcomestudien; Pharmakoepidemiologie, Gesundheitsökonomie; Endpunkte in Registerbasierten RCTs

Die Antworten auf die Frage: „Welche konkreten Forschungsfrage(n) halten Sie für besonders geeignet um sie mittels Verknüpfung mit GKV-Routinedaten von Krankenkassen im Kontext von NAPKON zu beforschen?“ spiegeln das breite Forschungsspektrum wider in dem mit GKV-Daten gearbeitet werden sollte.

Neben den bereits klassischen Forschungsinhalten von GKV-Routinedatennutzung (Arbeitsunfähigkeit, Inanspruchnahmemuster und deren Prädiktoren, Krankheitskostenermittlungen, Untersuchung von Versorgungspfaden und Leitlinienadhärenz) wurden insbesondere auch Datenverknüpfungsanwendungen genannt (z. B. Langzeitverlaufsuntersuchungen von Post-Covid Symptomen verknüpft mit AU-Zeiten, direkten und indirekten Krankheitskosten im Vergleich zu Propensity Score gematchten Kontrollen ohne COVID-19, Identifizierung von früheren Erkrankungen als Risikofaktoren für Post-Covid Symptome). Weitere Vorschläge waren vergleichende Längsschnittuntersuchungen mit gesunden Kohorten, risikofaktoren-basierte Fragestellungen, Mortalitätsstatistiken, die Untersuchung von seltenen Erkrankungen und Medikamenten-Interaktionen. Ein weiterer Punkt waren auch Studien zu Endpunkten in registerbasierten RCTs.

Frage 6: Haben Sie bereits Projekte mit Abrechnungsdaten von Krankenkassen durchgeführt? Und wenn ja, wie viele?

Nein (noch keine durchgeführt)	11
1 Projekt	1
2-5 Projekte	8
6-10 Projekte	2
> 10 Projekte	1

Etwa die Hälfte der BefragungsTeilnehmende hat bereits Erfahrung mit Abrechnungsdaten der Krankenkassen, meisten im Bereich von 2-5 Projekten.

Frage 7: Haben Sie bereits Projekte mit einer Verlinkung von Primärdaten und Daten aus weiteren Quellen durchgeführt? Und wenn ja, wie viele?

Nein (noch keine durchgeführt)	15
1 Projekt	6
2-5 Projekte	3
6-10 Projekte	1
> 10 Projekte	1

Ein kleinerer Teil der BefragungsTeilnehmende hat auch bereits Verlinkungen von primären Daten mit anderen Datenquellen durchgeführt - hier meist epidemiologische Studien und (Krebs)registerstudien.

Frage 8 Welche Datenquellen haben Sie dabei verwendet?

a)Epidemiologische Studien (z. B. Kohorten- oder Fall-Kontrollstudien))	9
b) Registerstudien	5
c)Krebsregister	2
d) Weitere , einzige Nennung: GKV-Routinedaten	4, davon GKV 1

Freitext in Frage 8:

Wenn möglich nennen Sie bitte die Studien / Datensammlungen/ Datenquellen, die verwendet wurden.

Primärdaten, Krankenversichertendaten, Krankenhausabrechnungsdaten

Studie läuft noch. Ziel: Auswirkung einer ärztlichen Zweitmeinung bei Knieendoprothese auf die Operationshäufigkeit und die weitere Versorgung.

KORA (Kooperative Gesundheitsforschung in der Region Augsburg) NAKO (Nationale Kohorte) SHIP (Study of Health in Pomerania)

Akutschmerzregister QUIPS mit BARMER GKV

Daten der Kassenärztlichen Vereinigung, Daten aus der Antibiotika-Resistenz-Surveillance am RKI Daten aus EARS-Net
GKV-Routinedaten INKAR-Daten
Projekte im Rahmen des Innovationsfonds (multiresistente Keime, teledermatologische KOnsile, Onkologie etc.)
SHIP, GANI_MED
AOK; KV; Popstudien (UKB, SHIP, KORA, ect)
Noch nicht publiziert
Leider nicht ganz klar, worauf sich die letzten beiden Fragen beziehen. NAPKON oder ob überhaupt schon Daten gesammelt wurden.

Frage 9: Bewerten Sie bitte das Potenzial von Sozialdaten für die Gesundheitsforschung allgemein:

	Sehr hoch (1)	(2)	(3)	(4)	Irrelevant (5)
	16	8	2	0	0

Das Potenzial von Sozialdaten wird im als meistens als sehr hoch bis hoch eingeschätzt (24/26).

Frage 10. Welche gesundheitsrelevanten Forschungsbereiche/Forschungsfragen haben Sie bereits mit Sozialdaten (GKV-Routinedaten) beforcht?

1	Langzeitanwendung von Opioiden in Deutschland inkl. Mortalität und Risikofaktoren für missbräuchliche Verwendung Krankheitskosten beim Fibromyalgiesyndrom und Reizdarmsyndrom
2	
3	Infektionskrankheiten (spez. Harnwegsinfekt)
4	Psychiatrie, onkologische Versorgungsforschung
5	Gebrechlichkeit/Frailty, multidimensional (Healthy aging and biological age)
6	Indikation zu Knieendoprothese
7	Inanspruchnahmeverhalten
8	keine
9	Evaluation von Selektivverträgen; Versorgungskosten von Patienten mit Rückenschmerzen; Versorgungsinanspruchnahme und -qualität von Patienten mit Depression; gesundheitsökonomische Evaluation von Influenza- und Pneumokokkenimpfungen; Krankheitslast und Versorgungskosten nach krankenhausbekannter Sepsis; Inanspruchnahme, Qualität und Kosten der Palliativversorgung im regionalen Vergleich; Inanspruchnahme medizinischer Leistungen durch Ärzte; Einsatz von Point-of-Care-Diagnostik bei Vertragsärzten
10	Akutschmerz
11	Morbidität, Arzneimittelverordnungen, Anzahl Krankenhausaufenthalte
12	Einsatz von Schnelltests in der ambulanten Versorgung
13	Projekte im Rahmen des Innovationsfonds (multiresistente Keime, teledermatologische KOnsile, Onkologie etc.) verschiedene Fragestellungen in Kooperation mit der AOK BW (Niereninsuffizienz, polymyalgia rheumatica etc.)
14	Risikofaktoren basierte Fragestellungen

15 Pharmakoepi

Das Potenzial von Sozialdaten wird von der Mehrheit der Befragten als sehr hoch oder hoch eingeschätzt (23/25). Das Potenzial der GKV-Routinedaten zeigt sich unter anderem in dem breiten Indikationsfeld, das die befragten Wissenschaftler nach eigenen Angaben bereits abdecken: Neben Infektionskrankheiten (v.a. Harnwegsinfektionen), Arzneimittelforschung ("Langzeitverwendung von Opioiden in Deutschland inkl. Mortalität und Risikofaktoren für Missbrauch"), Krankheitskosten (z. B. Fibromyalgie-Syndrom und Reizdarmsyndrom), wurden Fragen aus dem Bereich chronischer Erkrankungen ("Gebrechlichkeit", "Mehrdimensionales, gesundes Altern und biologisches Alter", Psychiatrie und onkologische Versorgungsforschung) bearbeitet. Entsprechend breit sind auch die Forschungsbereiche /Forschungsfragen, die die NUM Forschenden für besonders relevant halten, um sie zukünftig mittels GKV-Routinedaten zu beforschen: Neben allgemeineren Fragen zur Erforschung ganzer Indikationen ("Covid-19") oder bestimmter Sektoren des Gesundheitssystems (z. B. "Rehabilitation" - Einsatz von Heilmitteln und Einfluss auf Aktivität und Teilhabe/Arbeitsfähigkeit, insbesondere unter Berücksichtigung extrabudgetärer Diagnosen - Langzeitrehabilitation - wo werden Patienten wie versorgt? Welche Lücken gibt es? Rehabilitation im Sozialraum - wo fehlen Ansprechpartner im System?)" wurde Potenzial für Kostenstudien und Arzneimittelstudien identifiziert. Geeignet seien die GKV Routinedaten auch, um intersektorale und regional vergleichende Fragestellungen zu bearbeiten und die Wirkung von unterschiedlicher politisch-ökonomischer Rahmenbedingungen (z. B. Vergütungsstrukturen, Personalanforderungen) auf die Qualität und Kosten der Versorgung zu untersuchen.

Frage 11. Welche weiteren Forschungsbereiche/Forschungsfragen halten Sie für besonders relevant um sie zukünftig mittels GKV-Routinedaten zu beforschen??

1	Nutzung von Heilmitteln und Einfluss auf Aktivität und Teilhabe / Arbeitsfähigkeit besonders unter der Berücksichtigung extrabudgetärer Diagnosen Langzeitrehabilitation- wo werden die Pat wie versorgt, welche Lücken bestehen Rehabilitation im Sozialraum- wo fehlen Ansprechpartner im System
2	a) Direkte und indirekte Krankheitskosten bei chronischen Schmerzsyndromen b) Abbildung problematischer Medikationsverläufe (z. B. postoperative Dauertherapie mit Opioiden) oder missbräuchlicher Verwendung von cannabisbasierten Arzneimitteln und Identifikation von Risikofaktoren c) Polypharmazie, insbesondere Kombinationen von zentralwirksamen Medikamenten (z.- B. Opiode und Antidepressiva)
3	sehr viele (hier nur Beispiele) Multimorbidität/Polypharmazie Arzneimittelsicherheit Chronische Erkrankungen (längsschnittige Beobachtungsstudien mit langen follow-ups) ERhebung von Outcome-Daten bei Interventionsstudien (Versorgungsforschung, komplexe Interventionen)
4	Covid-19
5	Erforschung von Risikofaktoren
6	Versorgung bei Osteoporose
7	Langzeitoutcome, Einflussfaktoren vor Ereignis
8	Langzeitmorbidität z. B. post COVID
9	intersektorale Fragestellungen; regional vergleichende Fragestellungen; Wirkungen veränderter/unterschiedlicher Rahmenbedingungen (zB Vergütungsstrukturen,

	Personalanforderungen) auf die Qualität und Kosten der Versorgung; Versorgungsforschung zur Bewältigung des Fachkräftemangels und des demographischen Wandels
10	Hausärztliche Versorgung: Überweisungen, Verläufe, Kontakte zu anderen Spezialisten, Latenzzeiten, Diagnostik
11	Einsatz und Nutzen von Schnelltests in der ambulanten Versorgung Kosten-Effektivität von medizinischen Interventionen
12	Langzeitoutcome und Arbeitsfähigkeit nach Intensivtherapie (PICS) Verbindung zu SepNet (Sepsis Outcome) Auswirkungen von COVID-Infektions- und Impfzeitpunkt auf Krankheitsverlauf, insbesondere Komplikationen nach elektiven Eingriffen
13	perspektivisch seltene Erkrankungen
14	Policyevaluationen, Gesundheitsökonomie; Endpunkte in Registerbasierten RCTs
15	Nutzung von Heilmitteln und Einfluss auf Aktivität und Teilhabe / Arbeitsfähigkeit besonders unter der Berücksichtigung extrabudgetärer Diagnosen Langzeitrehabilitation- wo werden die Pat wie versorgt, welche Lücken bestehen Rehabilitation im Sozialraum- wo fehlen Ansprechpartner im System

Barrieren

Mit Hilfe von fünfstufigen Likert-Skalen wurde nach der Wichtigkeit/Relevanz verschiedener Barrieren für die Initiierung neuer Projekte mit Krankenversicherungsdaten gefragt und welche dieser Hindernisse als am relevantesten für die Initiierung neuer Projekte angesehen wurden.

Für die meisten der genannten Barrieren wurde deren Bedeutung als sehr relevant oder relevant eingeordnet. Als besonders hinderlich wurden „Schwierigkeiten mit Zugang und Kontaktaufnahme zu Krankenkassen“, „Fehlende Kooperationsbereitschaft/ mangelnde Ressourcen von Krankenkassen“ und „Fehlendes eigenes technisches Know-how / datenwissenschaftliche Kompetenzen zur Verarbeitung der GKV Daten“ dargestellt.

**Wir möchten gerne Barrieren zur Nutzung von Krankenkassen identifizieren.
Frage 12. Welche der folgenden Barrieren halten Sie für die Initiierung von neuen Projekten für besonders relevant?**

		Selbst erlebt? (ja/Nein)	Sehr relevant (1)	(2)	(3)	(4)	Irrelevant (5)
a)	Schwierigkeiten mit Zugang und Kontaktaufnahme zu Krankenkassen	15/9	11	7	3	3	0
b)	Fehlende Kooperationsbereitschaft/ mangelnde Ressourcen von Krankenkassen	14/10	4	11	8	1	0
c)	Schwierigkeiten bei der Beantragung der Beratung durch Ethikkommissionen	10/14	4	4	5	7	4

d)	Ablehnung bzw. hohe Auflage bei der Beratung durch Ethikkommissionen	5/19	1	3	10	6	4
e)	Ablehnung bzw. Auflage bei der Bewertung durch Datenschutzbeauftragte	11/13	8	7	6	1	2
f)	Ablehnung bzw. Auflage bei der Bewertung von Anträgen nach § 75 bei den zuständigen Landesbehörden	4/20	2	6	12	3	1
g)	Fehlendes eigenes technisches Know-how / datenwissenschaftliche Kompetenzen zur Verarbeitung der GKV Daten	-	5	7	9	2	2

Frage 12 h) Wenn möglich nennen Sie bitte weitere Hürden/Barrieren bei der Nutzung von GKV Routinedaten und insbesondere bei deren Verknüpfung mit Individualdaten.

1	Kosten, lange Wartezeiten bis zum Datenfluss, retrospektive Daten
2	Verzugsdauer von Erhebung bis Übermittlung
3	Verfolgung bei Kassenwechsel
4	In multizentrischen Projekten sind mehrere Datenschutzbeauftragte von Forschungseinrichtungen + Krankenkassen zuständig, die Abstimmung kann dann langwierig sein
5	Methodische Beratung im gesamten Prozess der Planung, Durchführung und Auswertung der Studie
6	Einzelinteressen beteiligter Krankenkassen; unterschiedliche Konzepte der Zusammenarbeit, je nach kooperierender Krankenkasse; schlechter Zugang zu (validen) PKV-Daten
7	Zum Teil schlechte Qualität der (Primär-)datenquellen (Missings, Fehler, etc.)
8	Pseudonymisierung der Versicherten-ID

Weitere Hinweise auf Hürden/Barrieren bei der Nutzung von GKV-Routinedaten und insbesondere bei deren Verknüpfung mit Individualdaten waren Kosten, lange Wartezeiten bis zum Datenfluss, der retrospektive Charakter der Daten, die Verzögerungszeit von der Erhebung bis zur Übermittlung und die Nachverfolgung von Patienten bei einem Kassenwechsel. Fehlende methodische Beratung bei der Planung, Durchführung und Auswertung von Studien, die Einbindung vieler regulatorischer Akteure in multizentrische Studien sowie individuelle Interessen der beteiligten Krankenkassen und deren je nach kooperierender Krankenkasse unterschiedlichen Konzepte der Zusammenarbeit wurden ebenfalls als Hürden identifiziert. Auch der unzureichende Zugang zu (validen) PKV Daten wurde als Problem identifiziert.

Frage 13) Ein großer Vorteil der Krankenkassendaten ist es aus wissenschaftlicher Sicht, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung der Patientinnen und Patienten einen Langzeitverlauf abbilden können, sofern die Forschungsfragen diesen Zeitraum notwendig machen. Welcher zukünftige Zeitraum ab der gegebenen Einwilligung ist aus Ihrer Sicht hier angemessen und für Patientinnen und Patienten absehbar? Bitte beachten Sie, dass ein Rückzug der Einwilligung jederzeit möglich ist.

Häufigkeit der Nennungen: 12 mal ‚Lebenslang‘, 3 mal 10 Jahre, 8 mal 5 Jahre

Aus wissenschaftlicher Sicht ist ein großer Vorteil von Krankenkassendaten, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung des Patienten einen Langzeitverlauf abbilden können, wenn Forschungsfragen diesen Zeitraum erforderlich machen.

Unter der Prämisse, dass die Einwilligung jederzeit widerrufen werden kann, stellt sich die Frage, für welchen zukünftigen Zeitraum der Datenerhebung ab der erteilten Einwilligung die Patienten ihre Zustimmung geben könnten. Der kürzeste von den Wissenschaftlern angegebene angemessene Zeitraum war 5 Jahre (n=3), aber die Mehrheit der Wissenschaftler hielt auch einen lebenslangen Zeitraum für angemessen (n=12). Dreimal wurden 10 Jahre als angemessener Zeitraum genannt.

Frage 14) Wir möchten Sie abschließend bitten, Maßnahmen zu bewerten um die Nutzung von GKV-Routinedaten (und insbesondere deren Verknüpfung) in Forschungsprojekten zu erleichtern und zu beschleunigen					
Tabelle 2: Frage 14: Unterstützungsangebote	Sehr relevant (1)	(2)	(3)	(4)	Irrelevant (5)
Zentrale Aufbereitung und Bereitstellung von GKV Routinedaten	15	6	0	0	0
Zentrale Qualitätssicherung der Daten	15	5	1	0	0
Einheitliches zentrales Verfahren für regulatorischen Anforderungen (Ethik, Datenschutz, §75 Genehmigungsverfahren)	16	4	1	0	0
Bereichsübergreifende Regelung zur Einwilligung der Patienten zur Verarbeitung der Behandlungsdaten, sofern Patienten zustimmen (z. B. Broad Consent)	18	3	0	0	0

Zentrale Unterstützungsangebote wie die Aufbereitung und Bereitstellung von GKV Routinedaten sowie eine zentrale Qualitätssicherung der Daten wurde ebenso für sehr relevant gehalten, wie eine bereichsübergreifende Regelung zur Einwilligung der Patienten zur Verarbeitung der Behandlungsdaten (z. B. Broad Consent). Hohe bis sehr hohe Relevanz als Unterstützungsleistung hätte laut der befragten Wissenschaftlerinnen und Wissenschaftler

auch ein einheitliches zentrales Verfahren für regulatorischen Anforderungen (Ethik, Datenschutz, §75 Genehmigungsverfahren).

Darüber hinaus wurde die Gelegenheit genutzt, weitere Maßnahmen zu benennen, die die Nutzung der GKV-Routinedaten (und deren Verknüpfung) erleichtern könnten. Neben der Stärkung der interdisziplinären Zusammenarbeit zur Erhöhung der Qualität der Ergebnisse, wurde angeregt, zu klären, welche Zugriffsrechte existierten und wem die Daten (nicht) gehörten. Vorgeschlagen wurde die Einrichtung einer Vertrauensstelle sowie der Austausch mit Ländern, die bereits Erfahrungen mit Routinedaten haben. Auch die Verwendung eines einheitlichen Datenformats wurde vorgeschlagen.

In den abschließenden allgemeinen Hinweisen und Kommentierungen wurde der Wunsch geäußert, dass eine Verknüpfung mit Daten der Rentenversicherungsträger besonders für den Bereich der Rehabilitation zusätzlich wünschenswert wäre. Und dass es konkret für das FDZ auf die Umsetzungsmodalitäten ankommen an. Eine Zentralisierung dürfe nicht mit hohem administrativen Aufwand und zu langen Wartezeiten einhergehen. Die Kosten sollten überschaubar sein.

Frage 14 e) Wenn möglich, nennen Sie bitte weitere Maßnahmen, die eine Nutzung von GKV-Routinedaten und (und deren Verknüpfung) erleichtern können.

#	Antworttext:
1	Erleichterung interdisziplinärer Zusammenarbeit (die in Projekten bereits besteht) und somit qualitative Steigerung der Ergebnisse
2	Es müsste geklärt werden, welche Zugriffsrechte es gibt und wem die Daten (nicht) gehören. Einrichtung einer Treuhandstelle. Austausch mit anderen Ländern mit viel Erfahrung mit Routinedaten
3	Einheitliches Datenformat
4	Die wichtigsten Themen sind oben genannt
5	Methodische Beratung des Zugangs, Nutzung und Auswertung der Daten
6	öffentlich zugängliche, einheitliche Data Dictionaries; öffentlich zugängliche Bereitstellung von Standard-Prozeduren für typische Algorithmen zur Datenaufbereitung (Selektion von Versicherten; zeitliche Abgrenzung der Leistungsanspruchnahme in Bezug auf ein tagesgenaues Indexereignis etc.)
7	Schulungen durch die zentrale Stelle, die GKV-Daten zur Verfügung stellen wird
8	SOPs zur Nutzung und Auswertung
9	cloudbasiertes Trusted Research Environment

Frage 14 f) Hier haben Sie die Möglichkeit für weitere allgemeine Hinweise oder Kommentierungen:

1	Eine Verknüpfung mit Daten der Rentenversicherungsträger wäre besonders für den Bereich der Rehabilitation zusätzlich wünschenswert
2	Es kommt sehr auf die Umsetzung an. Eine Zentralisierung darf nicht mit hohem administrativen Aufwand und zu langen Wartezeiten einhergehen. Die Kosten sollten überschaubar sein.
3	Sehr wichtige Initiative, andere europäische Länder sind viel weiter

FRAGEBOGEN NUM Fach-und organspezifische Arbeitsgruppen (FOSA)

Sehr geehrte Kolleginnen und Kollegen,

beim Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) wir ein Forschungsdaten-Zentrum aufgebaut, in dem im ersten Schritt die pseudonymisierten Daten aller gesetzlich Krankenversicherten zusammengeführt und für die Forschung zugänglich werden sollen. In weiteren Ausbaustufen sollen auch weitere Datenquellen (z. B. aus der elektronischen Patientenakte) verknüpft werden.

Auf der Basis eines Informierten Einverständnisses der TeilnehmerInnen aus prospektiven Kohorten soll es möglich werden, deren Krankenkassendaten zu verknüpfen. Das BMG hat uns beauftragt, bis 15.12.2022 ein Gutachten zum Thema "Record Linkage der Daten des mit extern gehaltenen Forschungsdaten am Beispiel NAPKON" zu erarbeiten.

Empfehlungen des Gutachtens sollen in dem im Koalitionsvertrag angekündigten Gesundheitsdatennutzungsgesetz aufgegriffen werden.

Uns liegt es sehr am Herzen, in diesem Zusammenhang auch die Perspektive der Forschenden einfließen zu lassen.

Aus diesem Grund würden wir uns freuen, wenn Sie den beiliegenden kurzen Fragebogen beantworten und uns bis zum 11.11.2022 per E-Mail oder Fax zurückübersenden.

Sollten Sie Fragen oder weitere Anregungen hierzu haben können Sie gerne jederzeit Peter Ihle (peter.ihle@uk-koeln.de) oder Jens-Peter Reese (jens-peter.reese@uni-wuerzburg.de) kontaktieren.

Wir bedanken uns schon jetzt für Ihre Unterstützung und verbleiben

mit freundlichen Grüßen

Prof. Dr. Wolfgang Hoffmann (Universität Greifswald),
Dr. Olaf Schoffer (ZEGV, Med. Fakultät der TU Dresden),
Dr. Peter Ihle (Universität Köln)
Prof. Dr. Jens-Peter Reese (Med. Fakultät der Universität Würzburg)
Prof Dr. Jochen Schmitt (ZEGV, Med. Fakultät der TU Dresden)

EINLEITUNG

Um den akuten Verlauf der COVID-19-Krankheit zu erfassen, aber auch um deren langfristigen Folgen zu analysieren, wurden im Rahmen des NAPKON-Projekts drei Kohorten etabliert, die verschiedene Abschnitte des COVID-19 Krankheitsverlaufs abbilden können: NAPKON-POP (populationsbasierte Plattform), NAPKON-SÜP (sektorenübergreifende Plattform) und NAPKON-HAP (hochauflösende Plattform).

Diese individuell erhobenen klinischen Daten liefern präzise Informationen zu Diagnostik, Laborbefunden, Schweregradeinteilungen von COVID-19 und Symptomen sowie patient-reported outcome measures (PROMs).

In den NAPKON Kohorten ist über die Patienteninformation und informierte Einwilligungserklärung bereits die Verknüpfung dieser individuellen klinischen Daten mit den personenbezogenen GKV-Routinedaten der Krankenkassen angelegt. Die konsentbasierte personenbezogene Verlinkung dieser beiden Datentypen (Primärdaten und Routinedaten) ermöglicht eine weitaus umfassendere Analyse von Krankheits- und Versorgungsprozessen bei COVID-19 Infektionen als dies alleine mit den klinischen Daten möglich wäre.

GKV-Routinedaten liefern Diagnosen und Leistungen aus dem ambulanten und stationären Sektor sowie ambulante Arzneimitteldaten und ermöglichen damit die vollständige und sektorenübergreifende Abbildung von Krankheitsverläufen inkl. Angaben zu Vorerkrankungen sowie langfristigen Nachbeobachtungszeiträumen (bis zu mehreren Jahren). GKV-Routinedaten bilden eine umfassende, effizient nutzbare, qualitativ hochwertige Grundlage für die Untersuchung der

- Inanspruchnahme von Leistungen der Gesundheitsversorgung
- Risikofaktoren wie Vorerkrankungen / Komorbiditäten
- Pflegebedürftigkeit
- Versorgungswege, und der
- Wirksamkeit, Sicherheit und Kosten von medikamentösen Therapien und Versorgungskonzepten.

Einschränkend muss darauf hingewiesen werden, dass den Routinedaten die detaillierte klinische Beschreibung (wie beispielsweise zur Schwere von Erkrankungen) fehlen. Umwelt- und Verhaltensbedingte Einflussgrößen sind ebenfalls in GKV-Routinedaten nicht gut abbildbar.

Wenn Sie damit einverstanden sind, dass wir Sie als Beteiligte:n am Gutachten in der Danksagung (Acknowledgements) nennen dürfen, senden Sie bitte Ihre Kontaktdaten an:

Name, Vorname _____

E-Mail Adresse _____

Institution / Affiliation _____

Mit einer möglichen erneuten Kontaktaufnahme über die oben genannte E-Mail zum Zwecke einer Einladung zu einer vertieften Diskussion/ eines Workshops bin ich einverstanden.

- Ja
- Nein

Beginn der Befragung

Angaben zur Person

1. Welcher Fach- und Organspezifischer Arbeitsgruppe gehören Sie an?
-

2. Sind Sie

- Männlich
- weiblich
- divers

3. Welcher Altersgruppe gehören Sie an?

- <35 Jahre
- 36-40 Jahre
- 41-45 Jahre
- 46-50 Jahre
- 51-55 Jahre
- 56-60 Jahre
- >60 Jahre

Erfahrung mit Projekten

4. Befürworten Sie es, dass auf Basis der informierten Einwilligung durch Patientinnen und Patienten (Informed Consent) externe Daten (z. B. GKV-Routinedaten von Krankenkassen, Registerdaten und Befragungsdaten) individuell verknüpft und für die Gesundheitsforschung genutzt werden?
- ja
 - nein
5. Welche konkreten Forschungsfrage(n) halten Sie für besonders geeignet um sie mittels Verknüpfung mit GKV-Routinedaten von Krankenkassen im Kontext von NAPKON zu beforschen?

6. Haben Sie bereits Projekte mit Abrechnungsdaten von Krankenkassen durchgeführt? Und wenn ja, wie viele?
- Nein (noch keine durchgeführt)
 - 1 Projekt
 - 2-5 Projekte
 - 6-10 Projekte
 - > 10 Projekte
7. Haben Sie bereits Projekte mit einer Verlinkung von Primärdaten und Daten aus weiteren Quellen durchgeführt? Und wenn ja, wie viele?
- Nein (keine durchgeführt)
 - 1 Projekt
 - 2-5 Projekte
 - 6-10 Projekte
 - > 10 Projekte

8. Welche Datenquellen haben Sie dabei verwendet? (Mehrfachnennung möglich)

- Klinische Studien
- Epidemiologische Studien (z. B. Kohorten- oder Fall-Kontrollstudien)
- Registerstudien
- Krebsregister
- weitere
- _____
- _____
- _____

Wenn möglich nennen Sie bitte die Studien / Datensammlungen/ Datenquellen, die verwendet wurden.

9. Bewerten Sie bitte das Potenzial von Sozialdaten für die Gesundheitsforschung allgemein:

Sehr hoch irrelevant

10. Welche gesundheitsrelevanten Forschungsbereiche/Forschungsfragen haben Sie bereits mit Sozialdaten (GKV-Routinedaten) beforscht?

11. Welche weiteren Forschungsbereiche/Forschungsfragen halten Sie für besonders relevant um sie zukünftig mittels GKV-Routinedaten zu beforschen?

Barrieren

12. Wir möchten gerne Barrieren zur Nutzung von Krankenkassen identifizieren.

Welche der folgenden Barrieren halten Sie für die Initiierung von neuen Projekten für besonders relevant?

a) Schwierigkeiten mit Zugang und Kontaktaufnahme zu Krankenkassen

Sehr relevant irrelevant

Selbst schon erlebt? ja/nein

b) Fehlende Kooperationsbereitschaft/ mangelnde Ressourcen von Krankenkassen

Sehr relevant irrelevant

Selbst schon erlebt? ja/nein

c) Schwierigkeiten bei der Beantragung der Beratung durch Ethikkommissionen

Sehr relevant irrelevant

Selbst schon erlebt? ja/nein

d) Ablehnung bzw. hohe Auflage bei der Beratung durch Ethikkommissionen

Sehr relevant ○ ○ ○ ○ ○ ○ irrelevant

Selbst schon erlebt? ja/nein

e) Ablehnung bzw. Auflage bei der Bewertung durch Datenschutzbeauftragte

Sehr relevant ○ ○ ○ ○ ○ ○ irrelevant

Selbst schon erlebt? ja/nein

f) Ablehnung bzw. Auflage bei der Bewertung von Anträgen nach § 75 bei den zuständigen Landesbehörden

Sehr relevant ○ ○ ○ ○ ○ ○ irrelevant

Selbst schon erlebt? ja/nein

g) Fehlendes eigenes technisches Know-how / datenwissenschaftliche Kompetenzen zur Verarbeitung der GKV Daten

Sehr relevant ○ ○ ○ ○ ○ ○ irrelevant

Wenn möglich nennen Sie bitte weitere Hürden/Barrieren bei der Nutzung von GKV Routinedaten und insbesondere bei deren Verknüpfung mit Individualdaten.

13. Ein großer Vorteil der Krankenkassendaten ist es aus wissenschaftlicher Sicht, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung der Patientinnen und Patienten einen Langzeitverlauf abbilden können, sofern die Forschungsfragen diesen Zeitraum notwendig machen. Welcher zukünftige Zeitraum ab der gegebenen Einwilligung ist aus Ihrer Sicht hier angemessen und für Patientinnen und Patienten absehbar?

Bitte beachten Sie, dass ein Rückzug der Einwilligung jederzeit möglich ist.

- <=1 Jahr
- 5Jahre
- Lebenslang

Anderer Zeitraum_: _____

14. Wir möchten Sie abschließend bitten, Maßnahmen zu bewerten um die Nutzung von GKV-Routinedaten (und insbesondere deren Verknüpfung) in Forschungsprojekten zu erleichtern und zu beschleunigen

a) Zentrale Aufbereitung und Bereitstellung von GKV Routinedaten

Sehr hilfreich ----- ----- ----- ----- nicht hilfreich

b) Zentrale Qualitätssicherung der Daten

Sehr hilfreich ----- ----- ----- ----- nicht hilfreich

c) Einheitliches zentrales Verfahren für regulatorischen Anforderungen (Ethik, Datenschutz, §75 Genehmigungsverfahren)

Sehr hilfreich ----- ----- ----- ----- nicht hilfreich

d) Bereichsübergreifende Regelung zur Einwilligung der Patienten zur Verarbeitung der Behandlungsdaten, sofern Patienten zustimmen (z. B. Broad Consent)

Sehr hilfreich ----- ----- ----- ----- nicht hilfreich

Wenn möglich, nennen Sie bitte weitere Maßnahmen, die eine Nutzung von GKV-Routinedaten und (und deren Verknüpfung) erleichtern können.

Weitere Hinweise oder Kommentierungen:

8.5 Anlage 3: Anschreiben und Fragebogen Patientenvertretungen

Anschreiben zum Fragebogen

Sehr geehrte ...

die wissenschaftliche Nutzung von Daten aus der medizinischen Versorgung gewinnt in Deutschland und international zunehmend an Bedeutung. Vor diesem Hintergrund plant die Bundesregierung im Koalitionsvertrag die Einführung eines Gesundheitsdatennutzungsgesetzes.

Das Bundesministerium für Gesundheit (BMG) arbeitet aktuell intensiv an diesem Gesetzesvorhaben. In diesem Zusammenhang hat uns das BMG beauftragt, bis 15.12.2022 ein Gutachten zu erstellen. Wir sind eine Gruppe von 5 Wissenschaftlern unterschiedlicher deutscher Hochschulen mit Erfahrung in der Nutzung und Verknüpfung von Krankenkassendaten mit Befragungs- oder Untersuchungsdaten aus prospektiven Kohortenstudien. Bei unserem Gutachten soll es im Kern darum gehen, die notwendigen Rahmenbedingungen, Anforderungen und Umsetzungsmöglichkeiten für die wissenschaftliche Nutzung von zusammengeführten Daten der für die Forschung gewonnenen Daten mit den Routinedaten der Krankenkassen zu beschreiben. Dabei wird diese Zusammenführung nur für StudienTeilnehmende angestrebt, die der Zusammenführung nach ausführlicher Information ausdrücklich zugestimmt haben. Diese Möglichkeit der Zusammenführung dieser Daten auf Basis der Informierten Einwilligung ist wichtig, weil viele für die Verbesserung der Gesundheitsversorgung relevante Fragen häufig nur durch die Kombination von Befragungs- und Untersuchungsdaten mit Krankenkassendaten aussagefähige Schlussfolgerungen erlauben.

Im Rahmen des Gutachtens ist es uns und auch dem BMG besonders wichtig, die Perspektive der Patientinnen und Patienten einzubinden. Daher schreiben wir Sie heute an und bitten um Ihre Unterstützung. Im Anschluss an diesen erklärenden Text haben wir einen kurzen Fragebogen formuliert, den wir Sie bitten bis zum 20.11.22 auszufüllen und an Sekretariat.zegv@ukdd.de zurückzumailen. Ihre Angaben werden vertraulich behandelt und nur zusammengefasste Ergebnisse aller teilnehmender Patientenvertreter:innen werden im Gutachten berichtet.

In dem für das Gutachten herangezogenen Beispiel geht es um die Zusammenführung von Abrechnungsdaten der Krankenkassen mit Befragungs- und Untersuchungsdaten einer Patientenkohorte zu COVID-19 (NAPKON-Kohorte).

In der Populationsbasierten Plattform (NAPKON-POP) werden in geographisch definierten Sammelgebieten (Würzburg, Berlin und Kiel) Patient*innen mit überstandener akuter SARS-CoV-2-Infektion identifiziert und in ein diagnostisches Langzeitprogramm eingeladen. Es erfolgt eine Biomaterialsammlung (z. B. Blut) und die Erfassung der Langzeitfolgen der Covid-Infektion, Komorbiditäten sowie rückblickend eine detaillierte Erfassung des akuten Erkrankungsverlaufes. Eingeschlossen werden erwachsene Probanden, bei denen eine SARS-CoV2-Infektion mindestens 6 Monate zurückliegt und die schriftlich oder elektronisch eingewilligt haben. Die NAPKON-Studie ist u. a. für die Erforschung von möglichen Langzeitfolgen von COVID-19 wichtig.

In unserem Beispiel haben die an der NAPKON-Studie teilnehmenden Patientinnen und Patienten nach entsprechender Aufklärung schriftlich eingewilligt, dass ihre im Rahmen der NAPKON-Studie erhobenen Daten mit Daten ihrer Krankenkasse verbunden werden. Die zuständige Ethikkommission hat diesem Vorgehen zugestimmt.

Um das Beispiel für Sie anschaulich zu machen, haben wir im folgenden Kasten Auszüge aus der Patientinnen und Patienten-Information aus der NAPKON-Studie zusammengestellt

Gesundheitsdaten sind Daten aus Ihrer Krankenversicherung. In der NAPKON-Studie möchten wir diese Daten wissenschaftlich verwenden, um gesundheitsfördernde, aber auch schädliche Einflüsse auf Ihren Gesundheitsverlauf nach der SARS-CoV-2 Infektion festzustellen, die nicht mit medizinischen Tests oder telefonischen Befragungen nach ihrem Wohlbefinden erfassbar sind.

Durch die Verknüpfung der im Rahmen der NAPKON-Studie erhobenen Befragungsdaten mit Daten Ihrer Krankenversicherung sollen folgende Fragen beantwortet werden:

1. Welcher Krankheitsverlauf zieht eine längere Arbeitsunfähigkeit nach sich?
2. Welche gemeinsamen Vorerkrankungen haben Personen, die an Langzeitfolgen der COVID-19 Erkrankung leiden?
3. Welche Risikofaktoren bestehen für die Entwicklung von Langzeitfolgen der COVID-19 Erkrankung (Long-COVID)?
4. Welche ärztlichen Leistungen und Arzneimittel erhalten die Menschen im Krankheitsverlauf?

5. Wie beeinflussen diese Behandlungen den Verlauf einer COVID-19 Erkrankung?

Die Beantwortung dieser Fragen ist wichtig, um neue Erkenntnisse zur Prävention von Long-COVID Erkrankungen zu ermitteln.

Zur Beantwortung dieser Fragestellungen bitten wir Sie um Ihre Erlaubnis, die folgenden Gesundheitsdaten Ihrer Krankenversicherung anzufordern, um diese wissenschaftlich zu nutzen. Die Auswahl der Daten beschränkt sich auf den Beobachtungszeitraum 2 Jahre vor und 5 Jahre nach Ihrer Infektion und umfasst:

- Informationen zur ambulanten und stationären medizinischen Versorgung (Diagnosen, erhaltene Leistungen, Verordnungen)
- Arzneimittelverordnungen
- Heil- und Hilfsmittelverordnungen (nicht-ärztliche Leistungen wie Physiotherapie, Rollstühle etc.)
- bei Erwerbstätigen Angaben zur Arbeitsunfähigkeit (Dauer, Diagnosen)

Wir würden uns sehr freuen, wenn Sie vor dem Hintergrund der oben genannten Informationen den beiliegenden Fragebogen bis zum 20.11.22 beantworten und an uns zurückschicken könnten. Bei Fragen können Sie uns gern kontaktieren unter Jochen.Schmitt@ukdd.de

Vielen Dank!

(Grußformel)

Befürworten Sie es, dass auf Basis der informierten Einwilligung durch Patientinnen und Patienten (Informed Consent) NAPKON-Studiendaten mit Krankenkassendaten individuell verknüpft und für die Gesundheitsforschung genutzt werden?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Würden Sie in dem skizzierten Beispiel und in ähnlichen Fällen den Mitgliedern Ihrer Patientenorganisation zur Studienteilnahme raten?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sehen Sie es als Auftrag an die Politik, die Nutzung von Krankenkassendaten und deren Verbindung zu Studiendaten (z.B. NAPKON) zu ermöglichen (z.B. durch entsprechende Rahmenbedingungen)?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sehen Sie es als Auftrag an die Krankenkassen, ihre Daten für die Forschung verfügbar zu machen, wenn die Versicherten ausdrücklich eingewilligt haben?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sehen Sie es als Auftrag an die beteiligten Forscher:innen, bei vorliegender Einwilligung zur Datenverknüpfung diese auch tatsächlich durchzuführen?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Erwächst aus Ihrer Sicht durch die Einwilligung in die Datennutzung für die teilnehmenden Patient:innen ein (moralischer) Anspruch, dass die vorgesehene Datenverknüpfung auch tatsächlich realisiert wird?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Ein großer Vorteil der Krankenkassendaten ist es aus wissenschaftlicher Sicht, dass diese Daten auch in der Zukunft erhoben werden und bei entsprechender Einwilligung der Patient:innen einen Langzeitverlauf abbilden können, sofern die Forschungsfragen diesen Zeitraum notwendig machen. Welcher zukünftige Zeitraum sollte durch eine gegebene Einwilligung abgedeckt werden bzw. ist aus Ihrer Sicht hier angemessen und für Patient:innen absehbar?

<= 1 Jahr (1)	5 Jahre (2)	lebenslang (3)	Anderer Zeitraum: (4)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____

Kennen Sie Studien aus Deutschland, bei denen eine Verbindung von Krankenkassendaten mit anderen Daten erfolgreich verknüpft wurden?

Ja → wenn Ja, welche: _____

Nein

Kennen Sie Studien aus Deutschland, bei denen eine Verbindung von Krankenkassendaten mit anderen Daten vorgesehen war, aber nicht realisiert werden konnte?

Ja → wenn Ja, welche: _____

Nein

Bisher haben v.a. die gesetzlichen Krankenversicherungen die Versorgungsdaten ihrer Versicherten so gut standardisiert, dass sie grundsätzlich für die Forschung nutzbar sind. Plädieren sie dafür, dass auch Private Krankenkassen die Versorgungsdaten ihrer Versicherten z.B. im skizzierten Beispiel für die Forschung verfügbar machen sollten?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Halten Sie es für eine Benachteiligung, dass privat versicherte Personen ihre Daten derzeit nicht in gleichem Maße für die Forschung verfügbar machen können, wie gesetzlich versicherte Personen?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Sollten nach Abschluss einer Studie die Ergebnisse interessierten Studienteilnehmer:innen zur Verfügung gestellt werden?

...starke Zustimmung (1)	...eher Zustimmung (2)	... keine klare Meinung (3)	...eher Ablehnung (4)	...starke Ablehnung (5)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Möchten Sie über die Ergebnisse dieser Befragung informiert werden?

- Ja
- Nein
- Keine Angabe

Würden Sie im weiteren Verlauf noch einmal für eine Befragung oder für eine Teilnahme an einem Workshop zur Verfügung stehen?

- Ja
- Nein
- Keine Angabe

Haben Sie weitere Anregungen oder Kommentare?

Bitte senden Sie den ausgefüllten Fragebogen per E-Mail: Sekretariat.ZEGV@uniklinikum-dresden.de an uns zurück!

Vielen Dank, dass Sie sich für eine Teilnahme an unserer Befragung entschieden haben!